

# Higher order Influence Functions and Minimax Estimation of Nonlinear Functionals

(Running Title: **Higher Order Influence Functions**)

by

James Robins<sup>a</sup>, Lingling Li<sup>b</sup>, Eric Tchetgen Tchetgen<sup>a</sup>, Aad van der Vaart<sup>c</sup>

<sup>a</sup>Harvard School of Public Health, <sup>b</sup>Harvard Pilgrim Health Care Institute,

<sup>c</sup>University of Leiden

## Abstract

Robins et al, 2008, published a theory of higher order influence functions for inference in semi- and non-parametric models. This paper is a comprehensive manuscript from which Robins et al, was drawn. The current paper includes many results and proofs that were not included in Robins et al due to space limitation. Particular results

---

**Key Words and Phrases:** U-statistics, Minimax, Influence Functions, Nonparametric, Robust Inference

**AMS 1991 Subject Classifications. Primary:** 123; **Secondary:** 456.

contained in the present paper that were not reported in Robins et al include the following. Given a set of functionals and their corresponding higher order influence functions, we show how to derive the higher order influence function of their product. We apply this result to obtain higher order influence functions and associated estimators for the mean of a response  $Y$  subject to monotone missingness under missing at random. These results also apply to estimating the causal effect of a time dependent treatment on an outcome  $Y$  in the presence of time-varying confounding. Finally, we include an appendix that contains proofs for all theorems that were stated without proof in Robins et al, 2008. The initial part of the paper is closely related to Robins et al,, the latter parts differ.

Specifically, we present a theory of point and interval estimation for nonlinear functionals in parametric, semi-, and non-parametric models based on higher order influence functions (Robins [18], Sec. 9; Li et al. [10], Tchetgen et al, [24], Robins et al, [20]). Higher order influence functions are higher order U-statistics. Our theory extends the first order semiparametric theory of Bickel et al. [3] and van der Vaart [30] by incorporating the theory of higher order scores considered by Pfanzagl [13], Small and McLeish [23] and Lindsay and Waterman [9]. The theory reproduces many previous results, produces new non- $\sqrt{n}$  results, and opens up the ability to perform optimal non- $\sqrt{n}$  inference in complex high dimensional models. We present novel rate-optimal point and interval estimators for various functionals of central impor-

tance to biostatistics in settings in which estimation at the expected  $\sqrt{n}$  rate is not possible, owing to the curse of dimensionality. We also show that our higher order influence functions have a multi-robustness property that extends the double robustness property of first order influence functions described by Robins and Rotnitzky [19] and van der Laan and Robins [27].

## 1 Introduction

Robins et al, 2008, published a theory of higher order influence functions for inference in semi- and non-parametric models. This paper is a comprehensive manuscript from which Robins et al, was drawn. The current paper includes many results and proofs that were not included in Robins et al due to space limitation. Particular results contained in the present paper that were not reported in Robins et al include the following. Given a set of functionals and their corresponding higher order influence functions, we show how to derive the higher order influence function of their product. We apply this result to obtain higher order influence functions and associated estimators for the mean of a response  $Y$  subject to monotone missingness under missing at random. These results also apply to estimating the causal effect of a time dependent treatment on an outcome  $Y$  in the presence of time-varying confounding. Finally, we include an appendix that contains proofs for all theorems that were stated without proof in Robins et al, 2008. The initial part of the paper is closely related to Robins

et al., the latter parts differ.

We have developed a theory of point and interval estimation for nonlinear functionals  $\psi(F)$  in parametric, semi-, and non-parametric models based on higher order likelihood scores and influence functions that applies equally to both  $\sqrt{n}$  and non- $\sqrt{n}$  problems (Robins 2004, Sec. 9; Li et al, 2006, Tchetgen et al, 2006, Robins et al, 2007). The theory reproduces results previously obtained by the modern theory of non-parametric inference, produces many new non- $\sqrt{n}$  results, and most importantly opens up the ability to perform non- $\sqrt{n}$  inference in complex high dimensional models, such as models for the estimation of the causal effect of time varying treatments in the presence of time varying confounding and informative censoring. See Tchetgen et al. (2007) for examples of the latter.

Higher order influence functions are higher order U-statistics. Our theory extends the first order semiparametric theory of Bickel et al. (1993) and van der Vaart (1991) by incorporating the theory of higher order scores and Bhattacharyya bases considered by Pfanzagl (1990), Small and McLeish (1994) and Lindsay and Waterman (1996).

The purpose of this paper is to demonstrate the scope and flexibility of our methodology by deriving rate-optimal point and interval estimators for various functionals that are of central importance to biostatistics. We now describe some of these functionals. We suppose we observe i.i.d copies of a random vector  $O = (Y, A, X)$  with unknown distribution  $F$  on each of  $n$  study subjects. In this paper, we largely

study non-parametric models that place no restrictions on  $F$ , other than bounds on both the  $L_p$  norms and on the smoothness of certain density and conditional expectation functions. The variable  $X$  represents a random vector of baseline covariates such as age, height, weight, hematocrit, and laboratory measures of lung, renal, liver, brain, and heart function.  $X$  is assumed to have compact support and a density  $f_X(x)$  with respect to the Lebesgue measure in  $R^d$ , where, in typical applications,  $d$  is in the range 5 to 100.  $A$  is a binary treatment and  $Y$  is a response, higher values of which are desirable. Then, in the absence of confounding by additional unmeasured factors, the functional  $\psi(F) = E\{E[Y|A=1, X]\} - E\{E[Y|A=0, X]\}$  is the mean effect of treatment in the total study population. Our results for  $E\{E[Y|A=1, X]\} - E\{E[Y|A=0, X]\}$  follow from results for the functional  $\psi(F) = E\{E[Y|A=1, X]\}$  based on data  $(AY, A, X)$  rather than  $(Y, A, X)$ . If  $Y$  is missing for some study subjects, and  $A$  is now the indicator that takes the value 1 when  $Y$  is observed and zero otherwise, then the functional  $E\{E[Y|A=1, X]\}$  is the marginal mean of  $Y$  under the missing at random assumption that the probability  $P[A=0|X, Y] = P[A=0|X]$  that  $Y$  is missing does not depend on the unobserved  $Y$ .

Returning to data  $O = (Y, A, X)$ , the functional

$$\begin{aligned}\psi(F) &= E\{Cov(Y, A|X)\} / E[var\{A|X\}] \\ &= E[w(X)\{E[Y|A=1, X] - E[Y|A=0, X]\}]\end{aligned}$$

with  $w(X) = var\{A|X\} / E[var\{A|X\}]$  is the variance weighted average treatment

effect. Our results for  $E\{Cov(Y, A|X)\}/E[var\{A|X\}]$  are derived from results for the functionals  $\psi(F) = E\{Cov(Y, A|X)\}$  and  $\psi(F) = E[\{E(Y|X)\}^2]$ .

We note that Robins and van der Vaart's (2006) construction of an adaptive confidence set for a regression function  $E(Y|X = x)$  depended on being able to construct a confidence interval for  $\psi(F) = E[\{E(Y|X)\}^2]$ . They constructed an interval for  $E[\{E(Y|X)\}^2]$  when the marginal distribution of  $X$  was known. In this paper, we construct a confidence interval for  $E[\{E(Y|X)\}^2]$  when the marginal of  $X$  is unknown and, in Section 5, use it to obtain an adaptive confidence set for  $E(Y|X = x)$ .

The functional  $E\{Cov(Y, A|X)\}$  is the functional  $E\{var(Y|X)\}$  in the special case in which  $Y = A$  wp1. Minimax estimation of  $var(Y|X)$  has recently been discussed by Wang et al. (2006) and Cai et al. (2006) in the setting of non-random  $X$ .

The function  $\gamma(x) = E[Y|A = 1, X = x] - E[Y|A = 0, X = x]$  is the effect of treatment on the subgroup with  $X = x$ . It is important to estimate the function  $\gamma(x)$ , in addition to the average treatment effect in the total population, because treatment should be given, since beneficial, to those subjects with  $\gamma(x) > 0$  but withheld, since harmful, from subjects with  $\gamma(x) < 0$ . We show that one can obtain adaptive confidence sets for  $\gamma(x)$  if one can set confidence intervals for the functional  $\psi(F) = E[\gamma(X)^2]$ . We construct intervals for  $E[\gamma(X)^2]$  under the additional assumption that the data  $O = (Y, A, X)$  came from a randomized trial. In a randomized trial, in

contrast to an observational study, the randomization probabilities,  $P(A = 1|X) = E(A|X)$  are known by design. We plan to report confidence intervals for  $E[\gamma(X)^2]$  with  $E(A|X)$  unknown elsewhere.

All of the above functionals  $\psi(F)$  have a positive semiparametric information bound (SIB) and thus a (first order) efficient influence function with a finite variance. In fact all the functionals  $\psi(F)$  have efficient influence function

$$IF(b(F), p(F), \psi(F)) \equiv if(O, b(X, F), p(X, F), \psi(F)) \quad (1)$$

where  $b(x, F), p(x, F)$  are monotone functions of certain conditional expectations, and, for any  $b^*(x), p^*(x)$ ,

$$E_F[IF(b^*, p^*, \psi(F))] = E_F[h_1(O) \{b^*(X) - b(X; F)\} \{p^*(X) - p(X; F)\}]$$

where  $h_1(O)$  is a known function. We refer to functionals in our class as doubly-robust to indicate that  $IF(b(F), p(F), \psi(F))$  continues to have mean zero when either (but not both)  $p(F)$  is misspecified as  $p^*$  or  $b(F)$  is misspecified as  $b^*$ . The functions  $b(x, F), p(x, F), if(O, b(X, F), p(X, F), \psi(F))$ , and  $h_1(O)$  differ depending on the functional  $\psi(F)$  of interest.

As the functionals  $\psi(F)$  are all closely related, we shall use  $E\{Cov(Y, A|X)\}$  as a prototype in this introduction. For  $\psi(F) \equiv E\{Cov(Y, A|X)\}$ ,  $b(X; F) = E_F(Y|X)$ ,  $p(X; F) = E_F(A|X)$ ,

$$IF(b(F), p(F), \psi(F)) = \{Y - b(X; F)\} \{A - p(X; F)\} - \psi(F),$$

and  $h_1(O) \equiv 1$ .

Whenever a functional  $\psi(F)$  has a non-zero SIB, given sufficiently stringent bounds on  $L_p$  norms and on smoothness, it is possible to use the estimated first order influence function to construct regular estimators and honest asymptotic confidence intervals whose width shrinks at the usual parametric rate of  $n^{-1/2}$ . We recall that, by definition, regular estimators are  $n^{1/2}$ -consistent. When  $X$  is high dimensional, the apriori smoothness restrictions on  $p(X; F)$  and  $b(X; F)$  necessary for point or interval estimators of  $E\{Cov(Y, A|X)\}$  to achieve the parametric rate of  $n^{-1/2}$  are so severe as to be substantively implausible. As a consequence, we replace the usual approach based on first order influence functions by one based on higher order influence functions.

To provide quantitative results, we require a measure of the maximal possible complexity (e.g. smoothness) of  $p(\cdot; F)$  and  $b(\cdot; F)$  believed substantively plausible.

We use Hölder balls for concreteness, although our methods extend to other measures of complexity. A function  $h(\cdot)$  lies in the Hölder ball  $H(\beta, C)$ , with Hölder exponent  $\beta > 0$  and radius  $C > 0$ , if and only if  $h(\cdot)$  is bounded in supremum norm by  $C$  and all partial derivatives of  $h(x)$  up to order  $\lfloor \beta \rfloor$  exist, and all partial derivatives of order  $\lfloor \beta \rfloor$  are Lipschitz with exponent  $(\beta - \lfloor \beta \rfloor)$  and constant  $C$ . We make the assumption that  $b(\cdot, F)$ ,  $p(\cdot, F)$  lie in given Hölder balls  $H(\beta_b, C_b)$ ,  $H(\beta_p, C_p)$ . Furthermore, it turns out we must also make assumptions about the complexity of the



function  $g(X; F) \equiv E_F[h_1(O|X)]f_X(X)$ , which we take to lie in a given  $H(\beta_g, C_g)$ .

For  $\psi(F) = E\{Cov(Y, A|X)\}$ ,  $g(X; F) = f_X(X)$

Using higher order influence functions, we construct regular estimators and honest (i.e uniform over our model) asymptotic confidence intervals for functionals  $\psi(F)$  in our class whose width shrinks at the usual parametric rate of  $n^{-1/2}$  whenever  $\beta/d \equiv \frac{\beta_b + \beta_p}{2}/d > 1/4$  and  $\beta_g > 0$ . This result cannot be improved on, since even when  $g(x)$  is known apriori,  $\beta/d > 1/4$  is necessary for a regular estimator to exist.

When  $\beta/d \leq 1/4$  and  $g(x)$  is known apriori, we have shown using arguments similar to those of Birge and Massart (1995) that the minimax rate of convergence for an estimator and minimax rate of shrinkage of a confidence interval is  $n^{-\frac{4\beta/d}{4\beta/d+1}} \geq n^{-\frac{1}{2}}$ . When  $g(x)$  is unknown, we construct point and interval estimators with this same rate of  $n^{-\frac{4\beta/d}{4\beta/d+1}}$  whenever

$$\beta_g/d > \beta/d \frac{2(\Delta + 1)(1 - 4\beta/d)}{(\Delta + 2)(1 + 4\beta/d) - 4(\beta/d)(1 - 4\beta/d)(\Delta + 1)}, \quad (2)$$

where  $\Delta = \left\lfloor \frac{\beta_p}{\beta_b} - 1 \right\rfloor$ . For example if  $\Delta = 0$ ,  $\beta/d = 1/8$ , we require  $\beta_g/d$  exceed  $1/22$  to achieve the rate  $n^{-\frac{4\beta/d}{4\beta/d+1}}$ . When the previous inequality does not hold and  $\Delta = 0$ , we have constructed, in a yet unpublished paper, estimators that converge at rate

$$\log(n) n^{-\frac{1}{2} + \frac{\beta_g/d}{1+2\beta_g/d} \frac{(m^*+1)^2}{2\beta/d}}, \text{ with} \quad (3)$$

$$m^* \equiv \left\lceil \left( \left[ \frac{\beta}{d} \left( 4\frac{\beta}{d} + \left( 1 - 4\frac{\beta}{d} \right) \frac{1 + 2\beta_g/d}{\beta_g/d} \right) \right]^{1/2} - (1 + 2\beta/d) \right) \right\rceil.$$

We conjecture that this rate is minimax, possibly up to log factors. In this paper,

however, the estimators we construct are inefficient when the previous inequality fails to hold, converging at rates less than the conjectured minimax rate of Eq (3).

Let us return to the case where  $Y = A$  wp1. Then  $\psi(F) = E\{var(Y|X)\}$  and  $p(\cdot) = b(\cdot)$  so  $\Delta = 0$ . Now, for fixed  $\beta$ , Eq (3) converges to  $\log(n)n^{-2\beta/d}$  as  $\beta_g \rightarrow 0$ , which agrees (up to a log factor) with the minimax rate of  $n^{-2\beta/d}$  given by Wang et al. (2006) and Cai et al. (2006) under the semiparametric homoscedastic model  $var(Y|X) = \sigma^2$  with equal-spaced non-random  $X$ . This result might suggest that  $X$  being random rather than equal-spaced can result in faster rates of convergence only when the density of  $X$  has some smoothness, as quantified here by  $\beta_g > 0$ . But this suggestion is not correct. Recall that we obtained the rate  $\log(n)n^{-2\beta/d}$  for  $\psi(F) = E\{var(Y|X)\}$  as  $\beta_g \rightarrow 0$  under a non-parametric model. In section 4, we construct a simple estimator of  $\sigma^2$  under the homoscedastic model with  $X$  random with unknown density that, for  $\beta/d < 1/4$ ,  $\beta < 1$ , and without smoothness restrictions on  $f_X(x)$ , converges at the rate  $n^{-\frac{4\beta/d}{4\beta/d+1}}$ , which is faster than the equal-spaced non-random minimax rate of  $n^{-2\beta/d}$ .

The paper is organized as follows. In Section 2, we define the higher order (estimation) influence functions of a functional  $\psi(F)$  for  $F$  contained in a model  $\mathcal{M}$  and prove two fundamental theorems - the extended information equality theorem and the efficient estimation influence function theorem. Further, in the context of a parametric model whose dimension increases with sample size, we outline why es-

timators based on higher order influence can outperform those based on first order influence functions in high-dimensional models. In Section 3, we introduce the class of functionals we study in the remainder of the paper and describe their importance in biostatistics. The theory of section 2, however, is not directly applicable to these functionals because they have first order but not higher order influence functions. We show that higher order influence functions fail to exist precisely because the Dirac delta function is not an element of the Hilbert space  $L_2$  of square integrable functions. We describe two approaches to overcoming this difficulty. The first approach is based on approximating the Dirac delta function by a projection operator onto a subspace of  $L_2$  of dimension  $k(n)$ , where  $k(n)$  can be as large as the square of the sample size  $n$ . The second approach is based on approximating the functional  $\psi(F)$  by a truncated functional  $\tilde{\psi}_{k(n)}(F)$ . The truncated functional has influence functions of all orders, is equal to  $\psi(F)$  if either a  $k(n)$  dimensional working parametric model (with  $k(n) < n^2$ ) for the function  $b(\cdot)$  or the function  $p(\cdot)$  in Eq. (1) is correct, and remains close to  $\psi(F)$  even if both working models are misspecified. We then use higher order influence function based estimators of  $\tilde{\psi}_{k(n)}(F)$  as estimators of  $\psi(F)$ . These estimators  $\hat{\psi}_{m,k(n)}$  are asymptotically normal with variance and bias for  $\psi(F)$  depending both on the choice of the dimension  $k(n)$  of the working models and on the order  $m$  of the influence function of  $\tilde{\psi}_{k(n)}(F)$ . We show that these same estimators  $\hat{\psi}_{m,k(n)}$  can also be obtained under the approximate Dirac delta function approach. We derive the

optimal estimator  $\widehat{\psi}_{m_{opt}, k_{opt}(n)}(\beta_b, \beta_p, \beta_g)$  in the class as a function of the Hölder balls in which the functions  $b, p$ , and  $g$  are assumed to lie. Finally we conclude section 3 by showing that the estimators  $\widehat{\psi}_{m, k(n)}$  have a multi-robustness property that extends the double-robustness property of the first order influence function estimator  $\widehat{\psi}_1$ .

In Section 4, we consider whether the estimators  $\widehat{\psi}_{m_{opt}, k_{opt}(n)}(\beta_b, \beta_p, \beta_g)$  are rate-minimax. We show that whenever  $\beta/d \equiv \frac{\beta_b + \beta_p}{2}/d > 1/4$  and  $\beta_g > 0$ ,  $\widehat{\psi}_{m_{opt}, k_{opt}(n)}(\beta_b, \beta_p, \beta_g)$  is not only rate minimax but is semiparametric efficient. Further, by letting the order  $m = m(n)$  of the U-statistic depend on sample size, we construct a single estimator  $\widehat{\psi}_{m(n), k(n)}$  that is semiparametric efficient for all  $\beta/d > 1/4$  even when  $g(\cdot)$  cannot be estimated at an algebraic rate. We show, however, that when  $\beta/d < 1/4$ ,  $\widehat{\psi}_{m_{opt}, k_{opt}(n)}(\beta_b, \beta_p, \beta_g)$  does not in general converge at the minimax rate. In Section 4.1, however, we construct a new estimator  $\widehat{\psi}_{\mathcal{K}_J}^{eff}(\beta_g, \beta_b, \beta_p)$  that converges at the minimax rate of  $n^{-\frac{4\beta/d}{4\beta/d+1}}$  whenever Eq. (2) holds. In Section 5, we use the results obtained in earlier sections to construct adaptive confidence intervals for a regression function  $E[Y|X = x]$  when the marginal of  $X$  is unknown and for the treatment effect function and optimal treatment regime in a randomized clinical trial. In Section 6.1, we discuss how to obtain higher order U-statistic point estimators and confidence intervals for functionals  $\tau(F)$  that are implicitly defined as the solution to an equation  $\psi(\tau, F) = 0$ . In Section 6.2, we define higher order testing influence functions and efficient scores and describe their relationship to the higher order es-

timation influence functions and efficient influence functions of Section 2. Finally, in Section 6.3, we discuss the relationship between the higher order U-statistic point estimators of an implicitly defined functional  $\tau(F)$  and higher order testing influence functions.

Before proceeding, several additional comments are in order. In this paper, we investigate the asymptotic properties of our higher order U-statistic point and interval estimators. The reader is referred to Li et al (2006) for an investigation of the finite sample properties of our procedures through simulation. In addition, precise regularity conditions are sometimes omitted from both the statements and the proofs of various theorems. This reflects the fact that the goal of this paper is to provide a broad overview of our theory as it currently stands.

Different subject matter experts will clearly disagree as to the maximum possible complexity of  $p(x; F)$ ,  $b(x; F)$  and  $g(x; F)$ . Thus it is important to have methods that adapt to the actual smoothness of these functions. Elsewhere, we plan to provide point estimators that optimally adapt to unknown smoothness. In contrast to point estimators, however, for honest confidence intervals, the degree of possible adaption to unknown smoothness is small. Therefore we propose that an analyst should report a mapping from apriori smoothness assumptions encoded in the exponents and radii of Hölder balls (or in other measures of complexity) to the associated optimal  $1-\alpha$  honest confidence intervals proposed in this paper. Such a mapping is finally only useful if

substantive experts can approximately quantify their informal opinions concerning the shape and wiggleness of  $p, b,$  and  $g$  using the measure of complexity on offer by the analyst. It is an open question which, if any, complexity measure is suitable for this purpose.

Finally, most of our mathematical results concern rates of convergence. We offer only a few results on the constants in front of those rates. This is not because the constant is less important than the rate in predicting how a proposed procedure will perform in the moderate sized samples occurring in practice. Rather, at present, we do not possess the mathematical tools necessary to obtain useful results concerning constants. A more extended discussion of the issue is found in Section 3 of Li et al. (2006).

In the following, we use  $X_n \asymp Y_n$  to mean  $X_n = O_p(Y_n)$  and  $Y_n = O_p(X_n)$ ;  $X_n \sim Y_n$  to mean  $\frac{X_n}{Y_n} \xrightarrow{P} 1$ ; and  $X_n \gg Y_n$  ( $X_n \ll Y_n$ ) to respectively mean  $\frac{Y_n}{X_n} \xrightarrow{P} 0$  ( $\frac{X_n}{Y_n} \xrightarrow{P} 0$ ) as  $n \rightarrow \infty$ .

## 2 Theory of Higher Order Influence Functions

Given  $n$  i.i.d observations  $\mathbf{O} \equiv \mathbf{O}_n \equiv \{O_i, i = 1, \dots, n\}$  from a model

$$\mathcal{M}(\Theta) = \{F(\cdot; \theta), \theta \in \Theta\},$$

we consider inference on a nonlinear functional  $\psi(\theta)$ . In general,  $\psi(\theta)$  can be infinite dimensional but for now we only consider the one dimensional case. In the following all

quantities can depend on the sample size  $n$ , including the support of  $O$ , the parameter space  $\Theta$ , and the functional  $\psi(\theta)$ . We generally suppress the dependence on  $n$  in the notation. We will be particularly interested in models in which the parameter  $\theta$  is infinite dimensional and  $\theta, \Theta$ , and  $\psi(\cdot)$  do not depend on  $n$ . We also briefly discuss models in which subvectors of  $\theta$  are finite-dimensional parameters whose dimension  $k(n) = n^{1+\rho}$  increases as power  $1 + \rho$  (often  $\rho > 0$ ) of  $n$  and thus  $\theta_n, \Theta_n$ , and  $\psi_n(\cdot)$  depend on  $n$ .

Our first task is to define higher order influence functions. Before proceeding we recall some facts about  $U$ -statistics. Consider a function  $b_m(o_1, o_2, \dots, o_m) \equiv b(o_1, o_2, \dots, o_m)$  where we often suppress  $b$ 's subscript  $m$ . For integers  $i_1, i_2, \dots, i_m$  lying in  $\{1, \dots, n\}$ , we define

$$B_{m,i_1,\dots,i_m} \equiv b_m(O_{i_1}, O_{i_2}, \dots, O_{i_m}) \equiv b(O_{i_1}, O_{i_2}, \dots, O_{i_m}).$$

and

$$\mathbb{V}_n[b_m] \equiv \frac{(n-m)!}{n!} \sum_{i_1 \neq i_2 \dots \neq i_m} B_{m,i_1,\dots,i_m}.$$

In an abuse of notation, we will consider the following expressions to be equivalent

$$\mathbb{V}_n[B_m] \equiv \mathbb{V}_n[B_{m,i_1,\dots,i_m}] \equiv \mathbb{V}_n[b_m].$$

Thus  $\mathbb{V}_n[b_m]$  is a  $m$ th order  $U$ -statistic with kernel  $b_m(o_1, o_2, \dots, o_m)$ . We do not assume that  $b_m(o_1, o_2, \dots, o_m)$  is symmetric. We will write  $\mathbb{V}_n[B_m]$  as  $\mathbb{B}_{n,m}$ . So, suppressing the dependence on  $n$ ,  $\mathbb{B}_m \equiv \mathbb{V}[B_m]$ .

Any  $\mathbb{B}_m$  has a unique (up to permutation) decomposition  $\mathbb{B}_m = \sum_{s=1}^m \mathbb{D}_s^{(b)}(\theta)$  under any  $F(\cdot; \theta)$  as a sum of degenerate U-statistics  $\mathbb{D}_s^{(b)}(\theta)$ , where degeneracy of  $\mathbb{D}_s^{(b)}(\theta)$  means that  $D_s^{(b)}(\theta) = d_s^{(b)}(O_{i_1}, O_{i_2}, \dots, O_{i_s}; \theta)$  satisfies

$$E_\theta [d_s^{(b)}(o_{i_1}, \dots, o_{i_{l-1}}, O_{i_l}, o_{i_{l+1}}, \dots, o_{i_s}; \theta)] = 0, l = 1, \dots, s$$

where upper and lower case letters, respectively, denote random variables and their possible realizations.

Let  $\mathcal{U}_m(\theta)$  be the Hilbert space of all  $U$ -statistics of order  $m$  with mean zero and finite variance with inner product defined by covariances with respect to the  $n$ -fold product measure  $F^n(\cdot; \theta)$ . Note that any  $U$ -statistic  $\mathbb{B}_s$  of order  $s$ ,  $s < m$ , is also an  $m$ th order  $U$ -statistic with  $\mathbb{D}_l^{(b)}(\theta)$  identically zero for  $m \geq l > s$ .

Since any two degenerate  $U$ -statistics of different orders are uncorrelated, the  $\mathcal{U}_m(\theta)$ -Hilbert space projection of  $\mathbb{B}_m$  on  $\mathcal{U}_l(\theta)$  is  $\sum_{s=1}^l \mathbb{D}_s^{(b)}(\theta)$  for  $l < m$ . Thus a  $U$ -statistic  $\mathbb{B}_m$  is degenerate  $\Leftrightarrow \mathbb{B}_m = \mathbb{D}_m^{(b)}(\theta) \Leftrightarrow \Pi_\theta[\mathbb{B}_m | \mathcal{U}_{m-1}(\theta)] = 0 \Leftrightarrow \mathbb{B}_m \in \mathcal{U}_{m-1}(\theta)^{\perp_{m,\theta}}$ , where  $\Pi_\theta[\cdot | \cdot] \equiv \Pi_{\theta,m}[\cdot | \cdot]$  is the projection operator of the Hilbert space  $\mathcal{U}_m(\theta)$  (with the dependence on  $m$  suppressed when no ambiguity can arise) and, for any linear subspace  $\mathcal{R}$  of  $\mathcal{U}_m(\theta)$ ,  $\mathcal{R}^{\perp_{m,\theta}}$  is its orthocomplement in the Hilbert space  $\mathcal{U}_m(\theta)$ . Given any  $\mathbb{B}_m = \mathbb{V}[B_m]$ ,  $\mathbb{D}_m^{(b)}(\theta)$  is explicitly given by  $\mathbb{V}[d_{m,\theta}\{B_m\}]$  where



$d_{m,\theta}$  maps  $B_m \equiv b(O_{i_1}, O_{i_2}, \dots, O_{i_m})$  to

$$d_{m,\theta} \{B_m\} = b(O_{i_1}, O_{i_2}, \dots, O_{i_m}) \quad (4)$$

$$+ \sum_{t=0}^{m-1} (-1)^{m-t} \sum_{i_{r_1} \neq i_{r_2} \dots \neq i_{r_t}} E_{\theta} (b(O_{i_1}, O_{i_2}, \dots, O_{i_m}) | O_{i_{r_1}}, O_{i_{r_2}}, \dots, O_{i_{r_t}})$$

Given a function  $g(\zeta)$ ,  $\zeta \equiv \{\zeta_1, \dots, \zeta_r\}^T$ , define for  $m = 0, 1, 2, \dots$ ,

$$g_{\bar{l}_m}(\zeta) \equiv g_{\setminus l_1, \dots, l_m}(\zeta) \equiv \frac{\partial^m g(\zeta)}{\partial \zeta_{l_1} \dots \partial \zeta_{l_m}}$$

with  $l_s \in \{1, \dots, r\}$  where the  $\setminus$  symbol denotes differentiation by the variables occurring to its right and the overbar  $\bar{l}_m$  denotes the vector  $(l_1, \dots, l_m)$ . Given a sufficiently smooth  $r$ -dimensional parametric submodel  $\tilde{\theta}(\zeta)$  mapping  $\zeta \in R^r$  injectively into  $\Theta$ , define for  $\theta$  in the range of  $\tilde{\theta}(\cdot)$ ,  $\psi_{\bar{l}_m}(\theta) \equiv \left( \psi \circ \tilde{\theta} \right)_{\setminus l_1, \dots, l_m}(\zeta) |_{\zeta = \tilde{\theta}^{-1}\{\theta\}}$  and  $f_{\bar{l}_m}(\mathbf{O}_n; \theta) \equiv \left( f \circ \tilde{\theta} \right)_{\setminus l_1, \dots, l_m}(\zeta) |_{\zeta = \tilde{\theta}^{-1}\{\theta\}}$ , where  $f(\mathbf{O}_n; \theta) \equiv \prod_i f(O_i; \theta)$  is the density of  $\mathbf{O}_n$  wrt a dominating measure. That is  $\psi_{\bar{l}_m}(\theta)$  and  $f_{\bar{l}_m}(\mathbf{O}_n; \theta)$  are higher order derivatives of  $\psi(\cdot)$  and  $f(\mathbf{O}_n; \cdot)$  under a parametric submodel  $\tilde{\theta}(\zeta)$ , where the model  $\tilde{\theta}$  has been suppressed in the notation. An  $sth$  order score associated with the submodel  $\tilde{\theta}(\zeta)$  is defined to be

$$\tilde{\mathbb{S}}_{s, \bar{l}_s}(\theta) \equiv f_{\bar{l}_s}(\mathbf{O}_n; \theta) / f(\mathbf{O}_n; \theta)$$

where  $\tilde{\mathbb{S}}_{s, \bar{l}_s}(\theta)$  is a U-statistic of order  $s$ . To understand why  $\tilde{\mathbb{S}}_{s, \bar{l}_s}(\theta)$  is a  $U$ -statistic we provide formulae for an arbitrary score  $\tilde{\mathbb{S}}_{s, \bar{l}_s}(\theta)$  in terms of the subject specific scores

$$S_{l_1 \dots l_m, j}(\theta) \equiv f_{/l_1 \dots l_m, j}(O_j; \theta) / f_j(O_j; \theta)$$

$j = 1, \dots, n$  for  $s = 1, 2, 3$ . Suppressing the  $\theta$ -dependence, results in Waterman and Lindsay (1996) imply

$$\tilde{\mathbb{S}}_{1,\bar{l}_1} = \sum_j S_{l_1,j}$$

$$\tilde{\mathbb{S}}_{2,\bar{l}_2} = \sum_j S_{l_1 l_2,j} + \sum_{s \neq j} S_{l_1,j} S_{l_2,s}$$

$$\tilde{\mathbb{S}}_{3,\bar{l}_3} = \sum_j S_{l_1 l_2 l_3,j} + \sum_{s \neq j} S_{l_1 l_2,j} S_{l_3,s} + S_{l_3 l_2,j} S_{l_1,s} + S_{l_1 l_3,j} S_{l_2,s} + \sum_{s \neq j \neq t} S_{l_1,j} S_{l_2,s} S_{l_3,t}.$$

Note these equations express each  $\tilde{\mathbb{S}}_{m,\bar{l}_m}$  as a sum of degenerate U-statistics. We now define a  $m$ th order estimation influence function  $\mathbb{IF}_{m,\psi(\cdot)}(\theta) \equiv \mathbb{IF}_{m,\psi}(\theta) \equiv \mathbb{IF}_m(\theta)$  for  $\psi(\theta)$  where we suppress the dependence on  $\psi$  when no ambiguity will arise.

**Definition 1** A U-statistic  $\mathbb{IF}_m(\theta)$  of order  $m$  and finite variance is said to be an  $m$ th order estimation influence function for  $\psi(\theta)$  if (i)  $E_\theta[\mathbb{IF}_m(\theta)] = 0$ ,  $\theta \in \Theta$  and (ii) for  $s = 1, 2, \dots, m$  and every suitably smooth and regular (see Appendix)  $r$  dimensional parametric submodel  $\tilde{\theta}(\zeta)$ ,  $r = 1, 2, \dots, m$ ,

$$\psi_{\setminus \bar{l}_s}(\theta) = E_\theta \left[ \mathbb{IF}_m(\theta) \tilde{\mathbb{S}}_{s,\bar{l}_s}(\theta) \right].$$

Estimation influence functions need not always exist, but when they do they are useful for deriving point estimators of  $\psi$  with small bias and for deriving confidence interval estimators centered on an estimate of  $\psi$ . We will generally refer to estimation

influence functions simply as influence functions. We remark that  $\mathbb{IF}_m(\theta)$  is an influence function under the above definition if and only if it is one under the modified version in which the dimension of the parametric submodel  $\tilde{\theta}(\zeta)$  is unrestricted. A key result is the following theorem which is related to results of Small and McLeish (1994).

**Theorem 2 *Extended Information Equality Theorem:*** *Given a  $m$ th order influence function  $\mathbb{IF}_m(\theta)$ , for any smooth and regular submodels  $\tilde{\theta}(\zeta)$  and  $s \leq m$ ,*

$$\partial^s E_\theta \left[ \mathbb{IF}_m \left( \tilde{\theta}(\zeta) \right) \right] / \partial \zeta_{l_1} \dots \partial \zeta_{l_s} |_{\zeta = \tilde{\theta}^{-1}\{\theta\}} = -\psi_{\setminus \bar{l}_s}(\theta)$$

*Thus, if the functionals  $E_\theta [\mathbb{IF}_m(\theta^*)]$  and  $-\psi(\theta^*) - \psi(\theta)$  have bounded Fréchet derivatives w.r.t.  $\theta^*$  to order  $m+1$  for a norm  $\|\cdot\|$ ,*

$$E_\theta [\mathbb{IF}_m(\theta + \delta\theta)] = -[\psi(\theta + \delta\theta) - \psi(\theta)] + O(\|\delta\theta\|^{m+1})$$

*since the functions  $E_\theta [\mathbb{IF}_m(\theta^*)]$  and  $-\psi(\theta^*) - \psi(\theta)$  of  $\theta^*$  have the same Taylor expansion around  $\theta$  up to order  $m$ .*

The proof is in the Appendix. Define the  $m$ th order tangent space  $\Gamma_m(\theta)$  at  $\theta$  for the model  $\mathcal{M}(\Theta)$  to be the subspace of  $\mathcal{U}_m(\theta)$  formed by taking the closed linear span of all scores of order  $m$  or less as we vary over all regular parametric submodels  $\tilde{\theta}(\zeta)$  (whose range includes  $\theta$ ) of our model  $\mathcal{M}(\Theta)$ . We say a model is (locally) nonparametric for  $m$ th order inference if  $\Gamma_m(\theta) = \mathcal{U}_m(\theta)$ .

Given any  $m$ th order estimation influence function  $\mathbb{IF}_m(\theta)$ , define the  $m$ th order efficient estimation influence function to be

$$\mathbb{IF}_m^{eff}(\theta) = \Pi_\theta [\mathbb{IF}_m(\theta) | \Gamma_m(\theta)]$$

where  $\Pi_\theta [\cdot | \cdot] \equiv \Pi_{\theta, m} [\cdot | \cdot]$  is the  $\mathcal{U}_m(\theta)$  –projection operator. In the appendix, we prove the following:

**Theorem 3 *Efficient Estimation Influence Function Theorem* :**

1.  $\mathbb{IF}_m^{eff}(\theta)$  is unique in the sense that for any two  $m$ th order influence functions

$$\Pi_\theta [\mathbb{IF}_m^{(1)}(\theta) | \Gamma_m(\theta)] = \Pi_\theta [\mathbb{IF}_m^{(2)}(\theta) | \Gamma_m(\theta)] \quad a.s.$$

2.  $\mathbb{IF}_m^{eff}(\theta)$  is a  $m$ th order estimation influence function and has variance less than or equal to any other  $m$ th order estimation influence function.
3.  $\mathbb{IF}_m(\theta)$  is a  $m$ th order estimation influence function if and only if

$$\mathbb{IF}_m(\theta) \in \left\{ \mathbb{IF}_m^{eff}(\theta) + \mathbb{U}_m(\theta) ; \mathbb{U}_m(\theta) \in \Gamma_m^{\perp m, \theta}(\theta) \right\}$$

where  $\Gamma_m^{\perp m, \theta}(\theta)$  is the ortho-complement of  $\Gamma_m(\theta)$  in  $\mathcal{U}_m(\theta)$ .

4. If  $\mathbb{IF}_m(\theta)$  exists then  $\mathbb{IF}_s^{eff}(\theta)$  exists for  $s < m$  and  $\Pi_\theta [\mathbb{IF}_m(\theta) | \Gamma_s(\theta)] = \mathbb{IF}_s^{eff}(\theta)$ .
5. If the model  $\mathcal{M}(\Theta)$  is (locally) nonparametric, then

(a) there is at most one  $m$ th order estimation influence function  $\mathbb{IF}_m(\theta)$  for  $\psi(\theta)$ ,

(b)

$$\mathbb{IF}_m(\theta) = \mathbb{IF}_{m-1}(\theta) + \mathbb{IF}_{mm}(\theta)$$

where

$$\mathbb{IF}_{m-1}(\theta) = \Pi_{m,\theta} [\mathbb{IF}_m(\theta) | \mathcal{U}_{m-1}(\theta)]$$

and  $\mathbb{IF}_{mm}(\theta)$  is a degenerate  $m$ th order  $U$ -statistic and thus

$$E_\theta [\mathbb{IF}_{m-1}(\theta) \mathbb{IF}_{mm}(\theta)] = 0.$$

(c) (i): Suppose, for a given  $m \geq 2$ ,  $\mathbb{IF}_{m-1}(\theta)$  exists and a kernel

$if_{m-1,m-1}(o_{i_1}, \dots, o_{i_{m-1}}; \theta)$  of  $\mathbb{IF}_{m-1,m-1}(\theta)$  has a first order influence function with kernel  $if_{1,if_{m-1,m-1}(o_{i_1}, \dots, o_{i_{m-1}}; \cdot)}(O_{i_m}; \theta)$  for all  $o_{i_1}, \dots, o_{i_{m-1}}$  in a set  $\mathcal{O}_{m-1}$  which has probability 1 under  $F^{(m-1)}(\cdot, \theta)$ . Then  $\mathbb{IF}_m(\theta)$  exists and

$$m\mathbb{IF}_{m,m}(\theta) = \mathbb{V} \left( d_{m,\theta} \left[ if_{1,if_{m-1,m-1}(o_{i_1}, \dots, o_{i_{m-1}}; \cdot)}(O_{i_m}; \theta) \right] \right) \quad (5)$$

where the operator  $d_{m,\theta}$  is given in Eq. (4).

(ii) Conversely, if  $\mathbb{IF}_m$  exists then the symmetric kernel

$if_{m-1,m-1}^{sym}(o_{i_1}, \dots, o_{i_{m-1}}; \theta)$  of  $\mathbb{IF}_{m-1,m-1}(\theta)$  has a first order influence function for all  $o_{i_1}, \dots, o_{i_{m-1}}$  in a set  $\mathcal{O}_{m-1}$  which has probability 1 under  $F^{(m-1)}(\cdot, \theta)$ .

Further

$$m^{-1}d_{m,\theta} \left[ if_{1,if_{m-1,m-1}^{sym}(o_{i_1}, \dots, o_{i_{m-1}}; \cdot)}(O_{i_m}; \theta) \right] = if_{m,m}^{sym}(O_{i_1}, \dots, O_{i_m}; \theta).$$

**Remark 4** *Pfanzagl (1990) previously proved part 5.c(i) for  $m = 2$ . Our theorem offers a generalization of his result. Note, in part (i) of 5(c), we can always take the kernel to be the symmetric kernel.*

**Remark 5** *Provided one knows how to calculate first order influence functions, one can obtain  $\mathbb{IF}_2(\theta), \dots, \mathbb{IF}_m(\theta)$  recursively using part (5.c). An example of such a calculation is given in Section 3.2.2 below. Thus part (5.c) has the interesting implication that even though higher order influence functions are defined in terms of their inner products with higher order scores  $\tilde{S}_{m,\bar{l}_m}$ , nevertheless, in (locally) nonparametric models, one can derive all the higher order influence functions of a functional  $\psi(\theta)$  without even knowing how to compute the scores  $\tilde{S}_{m,\bar{l}_m}$  for any  $m > 1$ . In fact, one need not even be aware of the structure of the scores  $\tilde{S}_{m,\bar{l}_m}$  in terms of the subject-specific higher order scores  $S_{l_1 \dots l_s, j}(\theta)$ . In contrast, in parametric or semiparametric models whose tangent space  $\Gamma_m(\theta)$  does not equal the set  $\mathcal{U}_m(\theta)$  of all  $m$ th order  $U$ -statistics, one can often (but not always) still obtain an inefficient influence  $\mathbb{IF}_m(\theta)$  by applying part (5.c) of the Theorem. However, calculation of the efficient influence function  $\mathbb{IF}_m^{eff}(\theta) = \Pi_\theta[\mathbb{IF}_m(\theta) | \Gamma_m(\theta)]$  by projection generally requires explicit knowledge of the scores  $\tilde{S}_{m,\bar{l}_m}$  to derive  $\Gamma_m(\theta)$ . For this reason, it can be considerably more difficult to analyze certain parametric models (with dimension increasing with sample size) than to analyze (locally) nonparametric models. We will consider derivation of and projections onto  $\Gamma_m(\theta)$  in a forthcoming paper. In the current paper, however, we do*

calculate  $IF_2^{eff}(\theta)$  in one model that is not (locally) nonparametric so as to provide some sense of the issues that arise. Specifically in Section 4, we calculate  $IF_2^{eff}(\theta)$  for the functional  $E[\{E[Y|X]\}^2]$  in a model that assumes the marginal distribution of  $X$  is known.

**Remark 6 : Implications of Theorem (3) for the Variance of Unbiased**

**Estimators:** Suppose we have  $n$  iid draws  $\mathbf{O} = (O_1, \dots, O_n)$  from  $F(o; \theta), \theta \in \Theta$ , and a  $U$ -statistic  $\hat{\psi}_m$  of order  $m \leq n$  with  $\text{var}_\theta[\hat{\psi}_m] < \infty$  for  $\theta \in \Theta$  satisfying  $E_\theta[\hat{\psi}_m] = \psi(\theta)$  for all  $\theta \in \Theta$ . That is,  $\hat{\psi}_m$  is unbiased for  $\psi(\theta)$ . We will use Theorem (3) to generalize a number of well-known results on minimum variance unbiased estimation to arbitrary models.

By  $E_\theta[\hat{\psi}_m] = \psi(\theta)$ , we immediately conclude that, viewing  $\hat{\psi}_m$  as a  $k$ th order  $U$ -statistic,  $\hat{\psi}_m - \psi(\theta)$  is a  $k^{th}$  order estimation influence function for  $\psi(\theta)$  for  $n \geq k \geq m$ . By Theorem (3),  $\text{var}_\theta[\hat{\psi}_m] \geq \text{var}_\theta[\mathbb{IF}_m^{eff}(\theta)]$ . We refer to  $\text{var}_\theta[\mathbb{IF}_m^{eff}(\theta)]$  as the  $m$ th order Bhattacharyya variance bound at  $\theta$  for the parameter  $\psi(\theta)$  in model  $\mathcal{M}(\Theta)$ , as this definition, in a precise analogy to Bickel et al. (1993)'s generalization of the Cramer-Rao variance bound, generalizes Bhattacharyya's (1947) variance bound to arbitrary semi- and non- parametric models. Indeed our 1st order Bhattacharyya bound is precisely Bickel et al.'s (1993) generalization of the Cramer-Rao variance bound.

We shall refer to an  $m$ th order  $U$ -statistic estimator  $\hat{\psi}_m$  as  $m$ th order 'unbiased

locally efficient' at  $\theta^*$  for  $\psi(\theta)$  in model  $\mathcal{M}(\Theta)$  if it is unbiased for  $\psi(\theta)$  under the model with variance at  $\theta^*$  equal to the  $m$ th order Bhattacharyya bound at  $\theta^*$ . If  $\hat{\psi}_m$  is 'unbiased locally efficient' at  $\theta^*$  for all  $\theta^* \in \Theta$ , we say it is 'unbiased globally efficient'. By Theorem (3),  $\text{var}_\theta [\mathbb{IF}_k^{eff}(\theta)] \geq \text{var}_\theta [\mathbb{IF}_m^{eff}(\theta)]$  for  $n \geq k > m$ . As a consequence if an  $m$ th order 'unbiased locally efficient' estimator  $\hat{\psi}_{m,eff}$  exists at  $\theta^*$  then, for  $n \geq k \geq m$ ,  $\mathbb{IF}_k^{eff}(\theta^*) = \mathbb{IF}_m^{eff}(\theta^*)$  so the  $m$ th and  $k$ th order Bhattacharyya bounds are equal at  $\theta^*$  and  $\hat{\psi}_{m,eff}$  is also  $k$ th order 'unbiased locally efficient' at  $\theta^*$ .

From the fact that for, an unbiased estimator  $\hat{\psi}_m$ ,  $\hat{\psi}_m - \psi(\theta)$  is an  $m$ th order influence function, we conclude that the variance of  $\hat{\psi}_m$  attains the the bound  $\text{var}_{\theta^*} [\mathbb{IF}_m^{eff}(\theta^*)]$  at  $\theta^*$  if and only if  $\hat{\psi}_m - \psi(\theta^*) = \mathbb{IF}_m^{eff}(\theta^*)$ , It follows that  $\hat{\psi}_m$  is 'unbiased globally efficient' if and only if  $\hat{\psi}_m - \psi(\theta) = \mathbb{IF}_m^{eff}(\theta)$  for all  $\theta \in \Theta$ . We thus have proved the following theorem in the  $\Rightarrow$  direction. The  $\Leftarrow$  direction is immediate.

**Theorem 7 :** In a model  $\mathcal{M}(\Theta)$ , there exists an  $m$ th order unbiased globally efficient U-statistic estimator of  $\psi(\theta)$ , if and only if, for all  $\theta \in \Theta$ ,  $\mathbb{IF}_m^{eff}(\theta) + \psi(\theta)$  is a function  $\hat{\psi}_{m,eff}$  of the data  $\mathbf{O}$ , not depending on  $\theta$ . In that case,  $\hat{\psi}_{m,eff}$  is the unique unbiased globally efficient estimator.

In a locally nonparametric model all unbiased  $m$ th order estimators are unbiased globally efficient, as there is a unique  $m$ th order influence function. For example, the usual unbiased estimator  $\hat{\sigma}^2 = \sum_{i=1}^n \left\{ X_i - \sum_{j=1}^n X_j/n \right\}^2 / (n-1)$  of the variance of a random variable  $X$  is a second order U-statistic and thus is a  $k$ th order unbiased



globally efficient U-statistic for  $k \geq 2$  in the locally nonparametric model consisting of all distributions under which  $\hat{\sigma}^2$  has a finite variance.

In Section 4 we use the results from this remark to compare the relative efficiencies of competing rate-optimal unbiased estimators in a model which is not locally nonparametric.

We now describe the main heuristic idea behind using higher order influence functions. Technical details are suppressed. Consider the estimator

$$\hat{\psi}_m = \psi(\hat{\theta}) + \mathbb{IF}_m^{eff}(\hat{\theta}) \quad (6)$$

based on a sample size  $n$ , where  $\hat{\theta}$  is an initial rate optimal estimator of  $\theta$  from a separate independent training sample. That is we assume that our actual sample size is  $N$  and we randomly split the  $N$  observations into two samples: an analysis sample of size  $n$  and a training sample of size  $N - n$  where  $(N - n)/N = c^*$ ,  $1 > c^* > 0$ . We obtain our initial estimate  $\hat{\theta}$  from the training sample data. Sample splitting has no effect on optimal rates of convergence, although in the form described here does affect 'constants'. Throughout the paper, we derive the properties of our estimators conditional on the data in the training sample. In a later section, we describe how one can sometimes obtain an optimal constant by choosing  $(N - n)/N = N^{-\epsilon}$ ,  $\epsilon > 0$  rather than  $c^*$ .

**Remark 8** *Note that sample splitting is avoided in most statistical applications by using modern “empirical process theory” to prove that ‘plug-in’ estimators such as*

$\widehat{\psi}_m = \{\psi(\theta) + \mathbb{IF}_m^{eff}(\theta)\}_{\theta=\widehat{\theta}}$  that estimate  $\theta$  from the same sample used to calculate  $\mathbb{IF}_m^{eff}(\cdot)$  have nice statistical properties. However empirical process theory is not applicable in our setting because we are interested in function classes whose size (entropy) is so large that they fail to be Donsker. For this reason we initially believed that explicit sample splitting would be difficult to avoid in our methodology. However, in Robins et al. (2007), we describe a new method, more analogous to the jackknife than to sample splitting, that effectively allows one to use all the data for estimator construction.

Expanding and evaluating conditionally on the training sample (or equivalently on  $\widehat{\theta}$ ), we find by Theorem 2 that the conditional bias

$$E_{\theta} [\widehat{\psi}_m - \psi(\theta) | \widehat{\theta}] = \psi(\widehat{\theta}) - \psi(\theta) + E_{\theta} [\mathbb{IF}_m^{eff}(\widehat{\theta}) | \widehat{\theta}]$$

is  $O_p(\|\widehat{\theta} - \theta\|^{m+1})$  which decreases with  $m$  provided  $\|\widehat{\theta} - \theta\| < 1$ .

In theorem 28 below, we show that if

$$\sup_{o \in \mathcal{O}} |f(o; \widehat{\theta}) - f(o; \theta)| \rightarrow 0$$

as  $\|\widehat{\theta} - \theta\| \rightarrow 0$ , where  $f(o; \theta)$  is the density of  $O$  under  $\theta$  and  $\mathcal{O}$  has probability one under all  $\theta \in \Theta$ , then

$$var_{\theta} [\widehat{\psi}_m | \widehat{\theta}] \equiv var_{\theta} [\mathbb{IF}_m^{eff}(\widehat{\theta}) | \widehat{\theta}] = var_{\widehat{\theta}} [\mathbb{IF}_m^{eff}(\widehat{\theta})] (1 + O_p(\|\widehat{\theta} - \theta\|))$$

Now, by Theorem 3,  $var_{\hat{\theta}} \left[ \mathbb{FF}_m^{eff} \left( \hat{\theta} \right) \right]$  increases with  $m$ . Further,  $var_{\hat{\theta}} \left[ \mathbb{FF}_1^{eff} \left( \hat{\theta} \right) \right] \asymp 1/n$ , since, conditional on  $\hat{\theta}$ ,  $\mathbb{FF}_1^{eff} \left( \hat{\theta} \right)$  is the sample average of *iid* random variables.

To proceed further we shall need to be more explicit about the model  $\mathcal{M}(\Theta)$ . For now, we consider finite-dimensional parametric models whose dimension  $k(n)$  increases with sample size. That is  $\theta \equiv \theta_n$  depends on  $n$  and the dimension of  $\Theta \equiv \Theta_n$  is  $k(n)$ . Suppose  $k(n) \asymp n^\gamma, \gamma \geq 0$ . Let  $\hat{\theta}_n$  be the maximum likelihood estimator of  $\theta$ . If  $k(n)$  increases slower than the sample size (i.e.,  $\gamma < 1$ ), then, a) under regularity conditions,  $\|\hat{\theta}_n - \theta_n\| = O_p \left( \{k(n)/n\}^{1/2} \right) = O_p \left( n^{-\frac{1}{2}(1-\gamma)} \right)$  with  $\|\cdot\|$  the usual Euclidean norm in  $R^{k(n)}$ ; and b)  $var_{\hat{\theta}} \left[ \mathbb{FF}_m^{eff} \left( \hat{\theta} \right) \right]$ , although increasing with  $m$ , remains order  $1/n$ ; as a consequence, if  $m$  is chosen greater than the solution  $m^*$  to  $n^{-\frac{m^*+1}{2}(1-\gamma)} = n^{-1/2}$ , the bias of  $\hat{\psi}_m$  will be  $o_p \left( n^{-1/2} \right)$ , the rate of convergence will be the usual parametric rate of  $n^{-1/2}$ , and thus, for  $n$  sufficiently large, the squared bias of  $\hat{\psi}_m$  will be less than the variance. As a consequence, as discussed in section 3.2.5, we can construct honest (i.e uniform over  $\theta_n \in \Theta_n$ ) asymptotic confidence intervals centered at  $\hat{\psi}_{m^*}$  with width of order  $n^{-1/2}$ . Here is a concrete example.

**Example:** Suppose  $O = (Y, X)$  with  $Y$  Bernoulli and with  $X$  having a density with respect to the uniform measure  $\mu(\cdot)$  on the unit cube  $[0, 1]^d$  in  $R^d$ . Suppose  $\psi = E \left[ (E[Y|X])^2 \right]$ . Let  $\{z_l(\cdot)\} \equiv \{z_l(x); 1, 2, \dots\}$  be a countable, linearly independent, sequence of either spline, polynomial, or compact wavelet basis functions dense in

$L_2(\mu)$ . Set  $\bar{z}_k(x) = (z_1(x), \dots, z_k(x))^T$ . We assume

$$E(Y|X=x) \in \left\{ \begin{array}{c} b(x; \bar{\eta}_{k^*(n)}) \equiv \left[ 1 + \exp \left( -\bar{\eta}_{k^*(n)}^T \bar{z}_{k^*(n)}(x) \right) \right]^{-1}; \\ \bar{\eta}_{k^*(n)} \in \mathcal{N}_{k^*(n)} \end{array} \right\},$$

$$f_X(x) \in \left\{ \begin{array}{c} f(x; \bar{\omega}_{k^{**}(n)}) \equiv c(\bar{\omega}_{k^{**}(n)}) \exp \left[ \bar{\omega}_{k^{**}(n)}^T \bar{z}_{k^{**}(n)}(x) \right]; \\ \bar{\omega}_{k^{**}(n)} \in \mathcal{W}_{k^{**}(n)} \end{array} \right\}$$

where  $c(\bar{\omega}_{k^{**}(n)})$  is a normalizing constant and  $\mathcal{N}_{k^*(n)}$  and  $\mathcal{W}_{k^{**}(n)}$  are open bounded subsets of  $R^{k^*(n)}$  and  $R^{k^{**}(n)}$ . Hence,  $\Theta_n = \mathcal{N}_{k^*(n)} \times \mathcal{W}_{k^{**}(n)}$  has dimension  $k(n) = k^*(n) + k^{**}(n)$  and  $\psi(\theta) = \psi_n(\theta_n) = \int b^2(x; \bar{\eta}_{k^*(n)}) f(x; \bar{\omega}_{k^{**}(n)}) d\mu(x)$ .

He (2000) and Portnoy (1988) prove that, under regularity conditions,  $\|\hat{\theta}_n - \theta_n\| = O_p(\{k(n)/n\}^{1/2})$  when  $k(n) = n^\gamma \ll n$ . Below we shall see that  $\text{var}_{\hat{\theta}}[\mathbb{IF}_m^{eff}(\hat{\theta})|\hat{\theta}] \asymp 1/n$  for  $n^\gamma \ll n$ .

Consider next models whose dimension  $k(n) \asymp n^\gamma$  increases faster than  $n$  (i.e.,  $\gamma > 1$ ). In such models, the MLE  $\hat{\theta}_n$  is generally inconsistent and indeed there may exist no consistent estimator of  $\theta_n$ . In that case,  $\|\hat{\theta}_n - \theta_n\|$  fails to be  $o_p(1)$  and the conditional bias  $E_\theta[\hat{\psi}_m - \psi(\theta)|\hat{\theta}]$  may not decrease with  $m$ . In order to guarantee consistent estimators of  $\theta_n$  exist, it is necessary to place further apriori restrictions on the complexity of  $\Theta_n$ . Typical examples of complexity-reducing assumptions would be an  $\epsilon$ -sparseness assumption that only  $k(n)^\epsilon$ ,  $0 < \epsilon < 1$ , of the  $k(n)$  parameters are non-zero or a smoothness assumption that specifies that the rate of decrease of the  $j^{th}$  component of  $\theta_n$  is equal to  $1/j$  raised to a given (positive) power. Even

after imposing such complexity-reducing assumptions,  $\psi(\theta) \equiv \psi_n(\theta_n)$  may not be estimable at rate  $n^{-1/2}$ .

For instance consider the previous example but now with  $\gamma^*$  and  $\gamma^{**}$  exceeding 1, so  $k^{**}(n) = n^{\gamma^{**}} \gg n$ ,  $k^*(n) = n^{\gamma^*} \gg n$  and  $k(n) = k^{**}(n) + k^*(n) \asymp n^\gamma \gg n$  with  $\gamma = \max(\gamma^{**}, \gamma^*)$ . Consider the norms  $\|\bar{\eta}_{k^*(n)}\| = \{\int b^2(x; \bar{\eta}_{k^*(n)}) d\mu(x)\}^{1/2}$ ,  $\|\bar{\omega}_{k^{**}(n)}\|_p = \{\int f(x; \bar{\omega}_{k^{**}(n)})^p d\mu(x)\}^{1/p}$  and  $\|\theta\|_p = \|\bar{\eta}_{k^*(n)}\| + \|\bar{\omega}_{k^{**}(n)}\|_p$ . Suppose, under a particular smoothness assumption, optimal rate estimators  $\hat{\bar{\eta}}_{k^*(n)}$  and  $\hat{\bar{\omega}}_{k^{**}(n)}$  of  $\bar{\eta}_{k^*(n)}$  and  $\bar{\omega}_{k^{**}(n)}$  satisfy  $\|\hat{\bar{\eta}}_{k^*(n)} - \bar{\eta}_{k^*(n)}\| = O_p(n^{-\gamma_\eta})$  and  $\|\hat{\bar{\omega}}_{k^{**}(n)} - \bar{\omega}_{k^{**}(n)}\|_p = O_p(n^{-\gamma_\omega})$  for some  $\gamma_\eta > 0$ ,  $\gamma_\omega > 0$  and all  $p \geq 2$ . Hence,  $\|\hat{\theta} - \theta\|_p = O_p(\max\{n^{-\gamma_\eta}, n^{-\gamma_\omega}\})$ . For  $\gamma > 1$ , based on arguments given later, we expect that  $\text{var}_{\hat{\theta}}[\hat{\psi}_m - \psi(\theta) | \hat{\theta}] \asymp \frac{n^{(\gamma-1)(m-1)}}{n}$  and  $E_{\theta}[\hat{\psi}_m - \psi(\theta) | \hat{\theta}] = O_p\left(\|\hat{\bar{\eta}}_{k^*(n)} - \bar{\eta}_{k^*(n)}\|^2 \|\hat{\bar{\omega}}_{k^{**}(n)} - \bar{\omega}_{k^{**}(n)}\|_{m-1}^{m-1}\right) = O_p(n^{-2\gamma_\eta - (m-1)\gamma_\omega}) = O_p(\|\hat{\theta} - \theta\|_{m-1}^{m+1})$ .

To find the estimator  $\hat{\psi}_{m_{best}}$  in the class  $\hat{\psi}_m$  with optimal rate of convergence, let  $m^* = 1 + \frac{1-4\gamma_\eta}{(\gamma-1)+2\gamma_\omega}$  be the value of  $m$  that equates the order  $n^{-4\gamma_\eta - 2(m-1)\gamma_\omega}$  of the squared bias and the order  $\frac{n^{(\gamma-1)(m-1)}}{n}$  of the variance. Then  $m_{best} = \lfloor m^* \rfloor$  if the order  $n^{-4\gamma_\eta - 2(m-1)\gamma_\omega} + n^{(\gamma-1)(m-1)-1}$  of the mean squared error at  $\lfloor m^* \rfloor$  is less than or equal to that at  $\lceil m^* \rceil$ . Otherwise,  $m_{best} = \lceil m^* \rceil$ . The rate of convergence of  $\hat{\psi}_{m_{best}}$  will often be slower than  $n^{-1/2}$ . Note  $m_{best} = 1$  whenever  $\gamma > 2$ , regardless of  $\gamma_\eta$  and  $\gamma_\omega$ .

By using the estimator  $\hat{\psi}_{\lceil m^* \rceil}$  rather than  $\hat{\psi}_{m_{best}}$ , we can guarantee that the vari-

ance asymptotically dominates bias and construct honest (i.e uniform over  $\theta_n \in \Theta_n$ ) asymptotic confidence intervals centered at  $\widehat{\psi}_{[m^*]}$ . Of course, the sample size  $n$  at which, for all  $\theta_n \in \Theta_n$ , the finite sample coverage of the intervals discussed above is close to the asymptotic (i.e. nominal) coverage is generally unknown and could be very large. For this reason, a better, but unfortunately as yet technically out of reach, approach to confidence interval construction is discussed in section 3.2.5.

In contrast to the case of parametric models of increasing dimension, in the infinite dimensional models which we consider in the following section, the functionals  $\psi(\theta)$  of interest have first order influence functions  $\mathbb{IF}_1(\theta)$  but do not have higher order influence functions. As a consequence, an initial 'truncation' step is needed before we can apply the approach outlined in the preceding paragraph.

Finally, even in the case of parametric models with  $k(n) \gg n$  and complexity reducing assumptions imposed, , when the minimax rate for estimation of  $\psi(\theta)$  is slower than  $n^{-1/2}$ , the optimal estimator  $\widehat{\psi}_{m_{best}}$  in the class  $\widehat{\psi}_m$  will generally not be rate minimax. See Section 3.2.6 and Sections 4.1.1 for additional discussion.

### 3 Inference for a Class of Doubly Robust Functionals:

#### 3.1 The class of functionals:

In this Section we consider models in which the parameter  $\theta$  is infinite dimensional and  $\theta, \Theta$ , and  $\psi(\cdot)$  do not depend on  $n$ . We make the following 4 assumptions  $Ai) - Aiv)$ :

Ai) The data  $O$  includes a vector  $X$ , where, for all  $\theta \in \Theta$ , the distribution of  $X$  is supported on the unit cube  $[0, 1]^d$  ( or more generally a compact set) in  $R^d$  and has a density  $f(x)$  with respect to the Lebesgue measure. Further  $\Theta = \Theta_1 \times \Theta_2$  where  $\theta_1 \in \Theta_1$  governs the marginal law of  $X$  and  $\theta_2 \in \Theta_2$  governs the conditional distribution of  $O|X$ .

Aii ) The parameter  $\theta_2$  contains components  $b = b(\cdot)$  and  $p = p(\cdot)$ ,  $b : [0, 1]^d \rightarrow \mathcal{R}$  and  $p : [0, 1]^d \rightarrow \mathcal{R}$ , such that the functional  $\psi(\theta)$  of interest has a first order influence function  $\mathbb{IF}_{1,\psi}(\theta) = \mathbb{V} [IF_{1,\psi}(\theta)]$ , where

$$IF_{1,\psi}(\theta) = H(b, p) - \psi(\theta), \quad (7)$$

$$\text{with } H(b, p) \equiv h(O, b(X), p(X))$$

$$\equiv b(X)p(X)h_1(O) + b(X)h_2(O) + p(X)h_3(O) + h_4(O) \quad (8)$$

$$\equiv BPH_1 + BH_2 + PH_3 + H_4,$$

and the known functions  $h_1(\cdot), h_2(\cdot), h_3(\cdot), h_4(\cdot)$  do not depend on  $\theta$ .

Aiii )

a)  $\Theta_{2b} \times \Theta_{2p} \subseteq \Theta_2$  where  $\Theta_{2b}$  and  $\Theta_{2p}$  are the parameter spaces for the functions  $b$  and  $p$ . Furthermore the sets  $\Theta_{2b}$  and  $\Theta_{2p}$  are dense in  $L_2(F_X(x))$  at each  $\theta_1^* \in \Theta_1$ .

or

b)  $b^*(\cdot) = p^*(\cdot)$ ,  $h_3(O) = h_2(O)$  w.p.1, and  $\Theta_{2b} \subseteq \Theta_2$  is dense in  $L_2(F_X(x))$  at each  $\theta_1^* \in \Theta_1$ .

**Remark:** Aiiib) can be viewed as a special case of Aiiia) as discussed in Example 1a below, so we need only prove results under assumption Aiiia).

Assumptions Ai)-Aiii) have a number of important implications that we summarize in a Theorem and two Lemmas.

**Theorem 9** *Double-Robustness: Assume Ai)-Aiii) hold, and recall  $p$  and  $b$  are elements of  $\theta$ . Then*

$$E_{\theta} [H(b, p^*)] = E_{\theta} [H(b^*, p)] = E_{\theta} [H(b, p)] = \psi(\theta)$$

for all  $(p^*, b^*) \in \Theta_{2p} \times \Theta_{2b}, \theta \in \Theta$ .

**Proof.** :  $E_{\theta} [H(b^*, p)] - E_{\theta} [H(b, p)] = E_{\theta} [\{H_1 p(X) + H_2\} \{b(X) - b^*(X)\}]$  and  $E_{\theta} [H(b, p^*)] - E_{\theta} [H(b, p)] = E_{\theta} [\{H_1 b(X) + H_3\} \{p(X) - p^*(X)\}]$ . The theorem then follows from part 1) of the following lemma. ■

Theorem 9 states that  $H(\cdot, \cdot)$  has mean  $\psi(\theta)$  under  $F(\cdot; \theta)$  even when  $p$  is misspecified as  $p^*$  or  $b$  is misspecified as  $b^*$ . We refer to the functional  $\psi(\theta)$  as doubly robust because of this property. The next lemma shows that  $H(b^*, p^*)$  is not unbiased if both  $b$  and  $p$  are simultaneously misspecified. That is,  $E_{\theta} [H(b^*, p^*)] \neq \psi(\theta)$ .

**Lemma 10** *Assume Ai)-Aiii) hold. Then*

$$1. E_{\theta} [\{H_1 B + H_3\} | X] = E_{\theta} [\{H_1 P + H_2\} | X] = 0$$



$$2. E_{\theta} [H(b^*, p^*)] - E_{\theta} [H(b, p)] = E_{\theta} [(B - B^*)(P - P^*) H_1]$$

$$\text{and } \psi(\theta) \equiv E_{\theta} [H(b, p)] = E_{\theta} [-BPH_1 + H_4]$$

**Proof.** Part 1): By assumptions *Ai*) and *Aiiia*) we have paths  $\tilde{\theta}_l(t)$ ,  $l = 1, 2, \dots$ , in our model with  $\tilde{\theta}_l(0) = \theta$  and  $p_l(t) = p_l(x; t) = p(x) + tc_l(x)$ ,  $b_l(x; t) = b(x)$ ,  $F_l(x; t) = F(x)$  for  $l = 1, 2, \dots$ , where the sequence  $c_l(\cdot)$  is dense in  $L_2[F_X(x)]$ . Let  $S_l(\theta)$  be the score for path  $\tilde{\theta}_l(t)$  at  $t = 0$ . Then by  $\psi(\tilde{\theta}_l(t)) = E_{\tilde{\theta}_l(t)} [H(b, p_l(t))]$

$$\begin{aligned} d\psi(\tilde{\theta}_l(t)) / dt|_{t=0} &= E_{\theta} [\{H_1 B + H_3\} c_l(X)] \\ &+ E_{\theta} [H(b, p) S_l(\theta)] \end{aligned}$$

$$\text{By } \mathbb{IF}_{1,\psi}(\theta) = H(b, p) - \psi(\theta),$$

$$d\psi(\tilde{\theta}_l(t)) / dt|_{t=0} = E_{\theta} [H(b, p) S_l]$$

Thus  $E[\{H_1 B + H_3\} c_l(X)] = 0$ . But  $\{c_l(\cdot)\}$  is dense in  $L_2[F_0(X)]$  so

$$E[H_1 B + H_3 | X] = 0$$

An analogous argument proves  $E_{\theta} [\{H_1 P + H_2\} | X] = 0$ . Part 2):  $E_{\theta} [H(b^*, p^*)] -$

$$E_{\theta} [H(b, p)] =$$

$$\begin{aligned} &E_{\theta} [(B^* P^* - BP) H_1 + (B^* - B) H_2 + (P^* - P) H_3] \\ &= E_{\theta} [(B^* P^* - BP) H_1 - (B^* - B) PH_1 - (P^* - P) BH_1] \\ &= E_{\theta} [(B - B^*)(P - P^*) H_1] \end{aligned}$$

where the second equality is by part 1). Choosing  $P^* = B^* = 0$  wp1 completes the proof of the theorem since then  $E_\theta [H(b^*, p^*)] = E_\theta [H_4]$ . ■

Below we will need the following partial converse to Lemma 10.

**Lemma 11** *Let  $\Theta_{2b}, \Theta_{2p}, \Theta_1$  and  $\Theta$  and  $H(b, p)$  be as defined in Ai)- Aiiia). Suppose that*

$$E_\theta [\{H_1 B + H_3\} | X] = E_\theta [\{H_1 P + H_2\} | X] = 0$$

*and  $\psi(\theta) = E_\theta [H(b, p)]$ . Then  $\mathbb{V}[H(b, p) - \psi(\theta)]$  is the first order influence function of  $\psi(\theta)$ .*

**Proof.** : The influence function of the functional  $E_\theta [H(b^*, p^*)]$  for known functions  $b^*, p^*$  is  $\mathbb{V}[H(b^*, p^*) - E_\theta [H(b^*, p^*)]]$ . Thus by the linearity of first order influence functions, the Lemma is true if and only if for each  $\theta_0 \in \Theta$ , the functional  $\tau(b, p) = E_{\theta_0} [H(b, p)]$  with  $\theta_0$  fixed has influence function equal to 0 wp1 at  $(b, p) = (b_0, p_0) \subset \theta_0$ . That the influence function is equal to 0 follows from the fact that, under the assumptions of the Lemma, for sets  $\{c_l(\cdot)\}$  and  $\{d_l(\cdot)\}$  dense in  $L_2[F_0(X)]$ ,

$$\begin{aligned} & dE_{\theta_0} [H(b_0(X) + tc_l(X), p_0(X) + td_l(X))] / dt|_{t=0} \\ &= E_{\theta_0} [\{H_1 b_0(X) + H_3\} d_l(X)] + E_{\theta_0} [\{H_1 p_0(X) + H_2\} c_l(X)] = 0 \end{aligned}$$

■

Results of Ritov and Bickel (1990) and Robins and Ritov (1997) imply it is not possible to construct honest asymptotic confidence intervals for  $\psi(\theta)$  whose width

shrinks to 0 as  $n \rightarrow \infty$  if  $b(\cdot)$  and  $p(\cdot)$  are too rough. Therefore we also place apriori bounds on their roughness. Our bounds will be based on the following definition.

**Definition 12** *A function  $h(\cdot)$  with domain  $[0, 1]^d$  is said to belong to a Hölder ball  $H(\beta, C)$ , with Hölder exponent  $\beta > 0$  and radius  $C > 0$ , if and only if  $h(\cdot)$  is uniformly bounded by  $C$ , all partial derivatives of  $h(\cdot)$  up to order  $\lfloor \beta \rfloor$  exist and are bounded, and all partial derivatives  $\nabla^{\lfloor \beta \rfloor}$  of order  $\lfloor \beta \rfloor$  satisfy*

$$\sup_{x, x+\delta x \in [0, 1]^d} |\nabla^{\lfloor \beta \rfloor} h(x + \delta x) - \nabla^{\lfloor \beta \rfloor} h(x)| \leq C \|\delta x\|^{\beta - \lfloor \beta \rfloor}.$$

We note that the  $L_p, 2 < p < \infty$  and  $L_\infty$  rates of convergence for estimation of a marginal density or conditional expectation  $h(\cdot) \in H(\beta, C)$  are  $O\left(n^{-\frac{\beta}{2\beta+d}}\right)$  and  $O\left(\left(\frac{n}{\log n}\right)^{-\frac{\beta}{2\beta+d}}\right)$  respectively. We refer to an estimator attaining these rates as rate optimal.

Aiv) We assume  $b(\cdot)$ ,  $p(\cdot)$ , and  $g(\cdot)$  lie in given Hölder balls  $H(\beta_b, C_b)$ ,  $H(\beta_p, C_p)$ ,  $H(\beta_g, C_g)$  where

$$g(x) \equiv E\{H_1|X=x\}f(x) \tag{9}$$

Furthermore we assume  $g(X) > \sigma_g > 0$  wp1. Finally we assume, as can always be arranged by a suitable choice of estimator, that the initial training sample estimators  $\hat{b}(\cdot)$ ,  $\hat{p}(\cdot)$ , and  $\hat{g}(\cdot)$  are rate optimal, have more than  $\max\{\beta_b, \beta_g, \beta_p\}$  derivatives, and have  $L_\infty$  norm bounded by a constant  $c_\infty$ . Further  $\inf_{x \in [0, 1]^d} \hat{g}(x) > \sigma_g$ . The reason for the restrictions on  $g(\cdot)$  will become clear below.

The restrictions  $A_i - A_{iv}$  are the only restrictions common to all functionals and models in the class. Additional model and/or functional specific restrictions will be given below.

To motivate our interest in such a class of functionals and models we provide a number of examples. In each case, one can use Lemma (11) to verify that the influence function of  $\psi(\theta)$  is as given. All but examples 3 and 4 are examples of (locally) nonparametric models.

**Example 1:** Suppose  $O=(A, Y, X)$  with  $A$  and  $Y$  univariate random variables.

**Example 1a: Expected Product of Conditional Expectations:** Let  $\psi(\theta) = E_\theta[p(X)b(X)]$  where  $b(X) = E_\theta[Y|X]$ ,  $p(X) = E_\theta[A|X]$ . In this model

$$\begin{aligned} IF_{1,\psi}(\theta) &= p(X)b(X) - \psi(\theta) \\ &\quad + p(X)\{Y - b(X)\} + b(X)\{A - p(X)\} \end{aligned}$$

so  $H_1 = -1, H_2 = A, H_3 = Y, H_4 = 0$ .

We also consider the special case of this model where  $A = Y$  with probability one (*w.p.1*). Then, as in assumption *Aiii*b),  $b(X) = p(X)$  *w.p.1*,  $H_2 = H_3$  *wp1*. Then  $\psi(\theta) = E_\theta[b^2(X)]$ . In Section 5, we show how our confidence interval for  $E_\theta[b^2(X)]$  can be used to obtain an adaptive confidence interval for the regression function  $b(\cdot)$ .

**Example 1b : Expected Conditional Covariance**

$$\psi(\theta) = E_\theta[AY] - E_\theta[p(X)b(X)] = E_\theta[Cov_\theta\{Y, A|X\}]$$

has influence function

$$AY - \{p(X)b(X) + p(X)\{Y - b(X)\} + b(X)\{A - p(X)\}\} - \psi(\theta)$$

so  $H_1 = 1, H_2 = -A, H_3 = -Y, H_4 = AY$ .

The next example 1c shows that a confidence interval and point estimators for  $E_\theta[Cov_\theta\{Y, A|X\}]$  can be used to obtain confidence intervals and point estimator for the variance weighted average treatment effect in an observational study.

**Example 1c: Variance-weighted average treatment effect:** Suppose, in an observational study,  $O = \{Y^*, A, X\}$ ,  $A$  is a binary treatment taking values in  $\{0, 1\}$ ,  $Y^*$  is a univariate response and  $X$  is a vector of pretreatment covariates. Consider the parameter  $\tau(\theta)$  given by:

$$\tau(\theta) = \frac{E_\theta[cov_\theta(Y^*, A|X)]}{E_\theta[var_\theta(A|X)]} = \frac{E_\theta[cov_\theta(Y^*, A|X)]}{E_\theta[\pi(X)\{1 - \pi(X)\}]}, \quad (10)$$

where  $\pi(X) = pr(A = 1|X)$  is often referred to as the propensity score. We are interested in  $\tau(\theta)$  for several reasons. First, in the absence of confounding by unmeasured factors,  $\tau(\theta)$  is the variance-weighted average treatment effect since  $\tau(\theta)$  can be rewritten as  $E_\theta[w_\theta(X)\gamma(X; \theta)]$  where  $w_\theta(X) = \frac{var_\theta(A|X)}{E_\theta[var_\theta(A|X)]}$  and

$$\gamma(x; \theta) = E_\theta(Y^*|A = 1, X = x) - E_\theta(Y^*|A = 0, X = x)$$

is the average conditional treatment effect at level  $x$  of the covariates. Second, under the semiparametric model

$$\gamma(X; \theta) = v(\theta) \text{ w.p.1} \quad (11)$$

that assumes the treatment effect does not depend on  $X$ ,  $\tau(\theta) = v(\theta)$ . However since the model (11) may not hold and therefore the parameter  $v(\theta)$  may be undefined, we choose to make inference on  $\tau(\theta)$  without imposing (11).

Now if for any  $\tau \in R$ , we define  $\psi(\tau, \theta)$  to be

$$\psi(\tau, \theta) = E_{\theta}[\{Y^*(\tau) - E_{\theta}(Y^*(\tau)|X)\}\{A - E_{\theta}(A|X)\}]$$

with  $Y^*(\tau) = Y^* - \tau A$ , it is easy to verify that  $\tau(\theta)$  may also be characterized as the solution  $\tau = \tau(\theta)$  to the equation  $\psi(\tau, \theta) = 0$ . Thus inference on  $\tau(\theta)$  is easily obtained from inference on  $\psi(\tau, \theta)$ . In particular a  $(1-\alpha)$  confidence set for  $\tau(\theta)$  is the set of  $\tau$  such that a  $(1-\alpha)$  CI interval for  $\psi(\tau, \theta)$  contains 0. Therefore, with no loss of generality, we consider the construction of a  $(1-\alpha)$  CI for  $\psi(\tilde{\tau}, \theta)$  for a fixed value  $\tau = \tilde{\tau}$ , and write  $Y = Y^*(\tilde{\tau})$  and  $\psi(\theta) = \psi(\tilde{\tau}, \theta)$ . Thus  $\psi(\theta) = E_{\theta}[Cov_{\theta}\{Y, A|X\}]$  and we are in the setting of Example 1b.

In section 6, we show the rates at which the width of the confidence sets for  $\psi(\tilde{\tau}, \theta)$  and for  $\tau(\theta)$  shrink with  $n$  are equal.

**Example 2a: Missing at Random:** Suppose  $O = (AY, A, X)$  where  $Y$  is an outcome that is not always observed,  $A$  is the binary missingness indicator,  $X$  is a  $d$ -dimensional vector of always observed continuous covariates, and let  $b(X) = E(Y|A = 1, X)$ ,  $\pi(X) = P(A = 1|X)$  be the propensity score, and  $p(X) = 1/\pi(X)$ .

We suppose  $\pi(X) > \sigma > 0$  and define

$$\psi(\theta) = E_{\theta} \left[ \frac{AY}{\pi(X)} \right] = E_{\theta} [b(X)] \quad (12)$$

Interest in  $\psi(\theta)$  lies in the fact that  $\psi(\theta)$  is the marginal mean of  $Y$  under the missing (equivalently, coarsening) at random (MAR) assumption that  $P(A = 1|X, Y) = \pi(X)$ . In this model  $IF_{1,\psi}(\theta) = Ap(X)(Y - b(X)) + b(X) - \psi(\theta)$  so  $H_1 = -A, H_2 = 1, H_3 = AY, H_4 = 0$ .

Note that if one has assumed apriori that  $f_X(\cdot)$  and  $p(X)$  lay in Hölder balls with respective exponents  $\beta_{f_X}$  and  $\beta_p$ , then  $\beta_g$  would be  $\min(\beta_{f_X}, \beta_p)$ , since  $g(X) = -f_X(X)/p(X)$ .

**Example 2b: Missing Not-at Random:** Consider again the setting of Example 2a but we no longer assume MAR. Rather we assume

$$P(A = 1|X, Y) = \{1 + \exp\{-[\gamma(X) + \alpha Y]\}\}^{-1}$$

may depend on  $Y$ , where now  $\gamma(X)$  is an unknown function and  $\alpha$  is a known constant (to be later varied in a sensitivity analysis). In this case the marginal mean of  $Y$  is given by  $\psi(\theta) = E_{\theta}(AY[1 + \exp\{-[\gamma(X) + \alpha Y]\}])$ . Robins and Rotnitzky (2001) proved this model places no restrictions on  $F(o)$  and derived

$$IF_{1,\psi}(\theta) = A\{1 + \exp\{-\alpha Y\}p(X)\}\{Y - b(X)\} + b(X) - \psi(\theta)$$

where, now,

$$b(X) = E[Y \exp\{-\alpha Y\} | A = 1, X] / E[\exp\{-\alpha Y\} | A = 1, X]$$

and  $p(X) = \exp\{-\gamma(X)\}$ . Thus

$$H_1 = -\exp\{-\alpha Y\} A, \quad H_2 = \{1 - A\}, \quad H_3 = AY \exp\{-\alpha Y\},$$

and  $H_4 = AY$ . When  $\alpha = 0$  this provides an alternate parametrization of Example 2a.

### Example 3: Marginal Structural Models and The Average Treatment

**Effect:** Consider the set-up of Example 1c including the non-identifiable assumption of no unmeasured confounders, except now  $A$  is discrete with possibly many levels and  $f(a|X) > \delta > 0$  wp1. A marginal structural model assumes  $E_{f_X}\{E_\theta(Y^*|A = a, X)\} = d(a, v(\theta))$ , where  $d(a, v)$  is a known function and  $v(\theta)$  is an unknown vector parameter of dimension  $d^*$ . When  $A$  is dichotomous with  $a \in \{0, 1\}$  and  $d(a, v) = v_1 + v_2 a$ , then  $v_2(\theta)$  is the average treatment effect parameter. Let  $f^*(a)$  be any density with the same support as  $A$  and let  $s^*(a)$  be a  $d^*$ -vector function, both chosen by the analyst. Then  $v(\theta)$  is identified as the (assumed) unique value of  $v$  satisfying

$$\psi_v(\theta) \equiv E_\theta \left[ s(O, A, v) \frac{f^*(A)}{f(A|X)} \right] = 0,$$

where  $s(O, a, v) = \{Y^* - d(a, v)\} s^*(a)$ . Thus a  $(1 - \alpha)$  confidence set for  $v(\theta)$  is the set of vectors  $v$  such that a  $(1 - \alpha)$  CI for  $\psi_v(\theta)$  contains 0. Therefore, with no loss of generality, we consider the construction of a  $(1 - \alpha)$  CI for the  $d$ -vector functional  $\psi(\theta) \equiv \psi_{\tilde{v}}(\theta)$  for a fixed value  $\tilde{v}$  and define  $h(O, A) \equiv s(O, a, \tilde{v})$  and



$b(a, X) \equiv E_\theta [h(O, a) | A = a, X]$ . Then  $\psi_{\widehat{v}}(\theta)$  has influence function

$$IF_1(\theta) = \frac{f^*(A)}{f(A|X)} \{h(O, A) - b(A, X)\} + \int b(a, X) dF^*(a) - \psi(\theta).$$

Next define  $p(a, X) = 1/f(a|X)$ ,  $\psi(\theta, a) = E_{f_X}[b(a, X)]$ . Then  $IF_1(\theta)$  is the integral

$$\begin{aligned} IF_1(\theta) &= \int dF^*(a) IF_1(a, \theta), \\ IF_1(a, \theta) &= H_1(a)p(a, X)b(a, X) \\ &\quad + H_2(a)b(a, X) + H_3(a)p(a, X) - \psi(\theta, a), \\ H_1(a) &= -I(A=a), H_2(a) = 1, H_3(a) = I(A=a)h(O, a). \end{aligned}$$

It follows that  $IF_1(\theta)$  is an integral over  $a \in A$  of influence functions  $IF_1(a, \theta)$  for parameters  $\psi(\theta, a)$  in our class with  $H_4(a) = 0$ . Thus we can estimate  $\psi(\theta)$  by  $\int dF^*(a) \widehat{\psi}(a)$ , where  $\widehat{\psi}(a)$  is an estimator of  $\psi(\theta, a)$ . If the support of  $A$  is of greater cardinality than  $d^*$ , the model is not locally nonparametric. Different choices for  $s^*(a)$  and  $f^*(a)$  for which  $\{\partial/\partial v^T\} E_\theta \left[ s(O, A, v) \frac{f^*(A)}{f(A|X)} \right]$  is invertible may result in different influence functions. All yield the same rate of convergence, although the constants differ. See Remark 5 above. Extension of our methods to continuous  $A$  will be treated elsewhere.

**Example 4: Confidence Intervals for The Optimal Treatment Strategy in a Randomized Clinical Trial:** Consider a randomized clinical trial with data  $O = \{Y, Y^*, A, X\}$ ,  $A$  is a binary treatment taking values in  $\{0, 1\}$ ,  $Y^*$  and  $Y$  are

univariate responses,  $X$  is a vector of pretreatment covariates. In a randomized trial, the randomization probabilities  $\pi_0(X) = P(A = 1|X)$  are known by design. Let  $b(x) = E_\theta(Y^*|A = 1, X = x) - E_\theta(Y^*|A = 0, X = x)$  and  $p(x) = E_\theta(Y|A = 1, X = x) - E_\theta(Y|A = 0, X = x)$  be the average treatment effects at level  $X = x$  on  $Y^*$  and  $Y$ . We assume  $Y$  and  $Y^*$  have been coded so that positive treatment effects are desirable. Let  $\psi(\theta) = E[b(X)p(X)]$ . Because the model is not locally nonparametric there exists more than a single first order influence function. Indeed, for any given function  $c(\cdot)$ ,

$$IF_{1,\psi}(\theta, c) = b(X)p(X) - \psi(\theta) + [b(X)\{Y - Ap(X)\} + p(X)\{Y^* - Ab(X)\}] \\ \times \{A - \pi_0(X)\}\sigma_0^{-2}(X) + c(X)\{A - \pi_0(X)\}$$

with  $\sigma_0^2(X) = \pi_0(X)\{1 - \pi_0(X)\}$  is an influence function in our class [provided it is square integrable] with  $H_1 = 1 - 2A\{A - \pi_0(X)\}\sigma_0^{-2}(X)$ ,  $H_2 = \{A - \pi_0(X)\}\sigma_0^{-2}(X)Y$ ,  $H_3 = \{A - \pi_0(X)\}\sigma_0^{-2}(X)Y^*$ ,  $H_4 = c(X)\{A - \pi_0(X)\}$ . As  $c(\cdot)$  is varied, one obtains all first order influence functions. We do not discuss the efficient choice of  $c(\cdot)$  in this paper.

Our interest lies in the special case where  $Y = Y^*$  wp1 (so there is but one response of interest) and thus, as in assumption *Aiiiib*),  $b = p$ ,  $H_2 = H_3$  and we construct confidence interval for  $\psi(\theta) = E[b^2(X)]$ . In Section 5 we describe how we can use a confidence interval for  $\psi(\theta) = E[b^2(X)]$  to obtain confidence intervals for

the treatment effect function  $b(x)$  and, most importantly, for the optimal treatment strategy  $d_{opt}(x) = I[b(x) > 0]$  under which a subject with covariate value  $x$  is treated if and only if the treatment effect  $b(x)$  is positive ( i.e.,  $d_{opt}(x) = 1$ ).

## 3.2 Higher Order Influence Functions for Our Model :

### 3.2.1 Dirac Kernels, Truncation Bias, and A Truncated Parameter

In all of our examples the functions  $p(\cdot)$  and  $b(\cdot)$  are functions of conditional expectations given the continuous random variable  $X$ . It is well known that the associated point-evaluation functional  $p(x)$  and  $b(x)$  do not have first order influence functions. It then follows from part 5c of Theorem 3 and the dependence of  $\mathbb{IF}_{1,\psi}(\theta) = \mathbb{V}[if_{1,\psi}(O_{i_1}; \theta)]$  on  $b(\cdot)$  and  $p(\cdot)$  evaluated at the point  $X$  that, in none of our examples, does  $\psi(\theta)$  have a second (or higher) order influence function.

As a precise understanding of the reason for the nonexistence of higher order influence functions for  $\psi(\theta)$  is fundamental to our approach, we now use part 5c of Theorem 3 to prove that  $\mathbb{IF}_{2,\psi}(\theta)$  does not exist by showing that the functional  $if_{1,\psi}(o; \theta)$  does not have a first order influence function  $\mathbb{V}[if_{1,if_{1,\psi}(o;\cdot)}(O; \theta)]$ . In this proof, we do not assume that  $b(\cdot)$  and  $p(\cdot)$  are functions of conditional expectations. Rather we only assume that our functional satisfies Assumptions A(i-iv). Let  $F_X$  and  $f_X = f_X(\cdot)$  denote the marginal CDF and density of  $X$ .

Consider paths (parametric submodels)  $\tilde{\theta}_l(t)$  such that  $\tilde{\theta}_l(0) = \theta$  satisfying

$$p_l(t) \equiv p_l(x, t) \equiv p(x) + tc_l(x),$$

$$b_l(t) \equiv b_l(x, t) \equiv b(x) + ta_l(x),$$

where the sequences  $c_l(\cdot)$  and  $a_l(\cdot)$ ,  $l = 1, 2, \dots$ , are each dense in  $L_2[F_X(x)]$ . Let

$$s_l(O; \theta) = s_l(O|X; \theta) + s_l(X; \theta),$$

$s_l(O|X; \theta)$ , and  $s_l(X; \theta)$  denote the overall, conditional, and marginal scores

$$\partial \ln f(O; \tilde{\theta}_l(0)) / \partial t, \partial \ln f(O|X; \tilde{\theta}_l(0)) / \partial t, \partial \ln f_X(X; \tilde{\theta}_l(0)) / \partial t$$

By linearity,  $if_{1,\psi}(o; \theta)$  has an influence function only if the functionals  $b(x)$  and  $p(x)$  have one as well. Now by differentiating the identity

$$E_{\tilde{\theta}_l(t)}[\{H_1 b_l(X, t) + H_3\} | X = x] = 0$$

wrt to  $t$  and evaluating at  $t = 0$ , we have

$$-E_\theta[\{\{H_1 b(X) + H_3\}\} s_l(O|X) | X = x] = E_\theta[H_1 | X = x] a_l(x)$$

However, by definition,  $b(x)$  has an influence function  $\mathbb{V}[if_{1,b(x)}(O; \theta)]$  at  $\theta$  only if for  $l = 1, 2, \dots$ , both  $\partial b_l(x, t) / \partial t|_{t=0} = a_l(x)$  equals  $E_\theta[if_{1,b(x)}(O; \theta) s_l(O; \theta)]$  and  $E_\theta[if_{1,b(x)}(O; \theta)] = 0$ . Thus if  $if_{1,b(x)}(O; \theta)$  exists, it must satisfy

$$\begin{aligned} & -E_\theta[\{H_1 b(X) + H_3\} s_l(O|X) | X = x] \\ & = E_\theta[H_1 | X = x] E_\theta[if_{1,b(x)}(O; \theta) s_l(O; \theta)] \end{aligned}$$

Without loss of generality, suppose  $H_1 \geq 0$  wp 1. Now if we could find a 'kernel'  $K_{f_X, \infty}(x, X)$  such that

$$\begin{aligned} r(x) &= E_{f_X} [K_{f_X, \infty}(x, X) r(X)] \\ &\equiv \int K_{f_X, \infty}(x, x^*) r(x^*) f_X(x^*) dx^* \text{ for all } r(\cdot) \in L_2(F_X) \end{aligned} \quad (13)$$

then

$$if_{1, b(x)}(O; \theta) \equiv - \left[ \begin{array}{l} \{E_\theta[H_1|X=x]\}^{-1/2} K_{f_X, \infty}(x, X) \\ \times \{E_\theta[H_1|X]\}^{-1/2} \{H_1 b(X) + H_3\} \end{array} \right]$$

would be an influence function since

$$\begin{aligned} &E_\theta[H_1|X=x] E_\theta \left[ \begin{array}{l} -\{E_\theta[H_1|X=x]\}^{-1/2} K_{f_X, \infty}(x, X) \times \\ \{E_\theta[H_1|X]\}^{-1/2} \{H_1 b(X) + H_3\} s_l(O; \theta) \end{array} \right] \\ &= E[H_1|X=x]^{1/2} E_\theta \left[ \begin{array}{l} -K_{f_X, \infty}(x, X) \{E_\theta[H_1|X]\}^{-1/2} \\ \times \{H_1 b(X) + H_3\} \{s_l(O|X) + s_l(X)\} \end{array} \right] \\ &= E[H_1|X=x]^{1/2} E_{f_X} \left\{ E_\theta \left[ \begin{array}{l} -K_{f_X, \infty}(x, X) \{E_\theta[H_1|X]\}^{-1/2} \times \\ \{H_1 b(X) + H_3\} s_l(O|X) | X \end{array} \right] \right\} \\ &= -E_\theta[(H_1 b(X) + H_3) s_l(O|X) | X=x] \end{aligned}$$

By an analogous argument

$$if_{1, p(x)}(O; \theta) = - \left[ \begin{array}{l} \{E_\theta[H_1|X=x]\}^{-1/2} K_{f_X, \infty}(x, X) \\ \times \{E_\theta[H_1|X]\}^{-1/2} \{H_1 p(X) + H_2\} \end{array} \right]$$

would be an influence function.

Indeed since the sequences  $\{c_l(\cdot)\}$  and  $\{a_l(\cdot)\}$  are dense the existence of such a kernel is also a necessary condition for  $if_{1,b(x)}(O;\theta)$  and  $if_{1,p(x)}(O;\theta)$  to exist and thus for  $if_{1,\psi}(o;\theta)$  to exist. A kernel satisfying Eq.(13) is referred to as the Dirac delta function wrt to the measure  $dF_X(x)$  and would clearly have to satisfy

$$K_{f_X,\infty}(x_{i_1}, x_{i_2}) = 0 \text{ if } x_{i_2} \neq x_{i_1} \quad (14)$$

were it to exist. Of course a kernel satisfying Eq. (13) is known not to exist in  $L_2[F_X] \times L_2[F_X]$ . We conclude that  $if_{1,\psi}(o;\theta)$  does not have an influence function and therefore  $\mathbb{IF}_{2,2,\psi}(\theta)$  does not exist.

**A Formal Approach:** To motivate how one might overcome this difficulty, we note that kernels satisfying Eq. (13) exist as generalized functions or kernels (also known as Schwartz functions or distributions). We shall "formally" derive higher order influence functions that appear to be elements of the space of generalized functions. However, we use these calculations only as motivation for statistical procedures based on ordinary kernels living in  $L_2[F_X] \times L_2[F_X]$ . Thus it does not matter whether these formal calculations could be made rigorous with appropriate redefinitions. Rather we can simply regard the following as results obtained by applying a "formal calculus" to part 5c of Theorem 3 that adds to the usual calculus additional identities licensed by Eqs. (13) and (14).

We will need the fact that, for any function  $v(x;\theta)$ , Eq. (14) implies that

$$v(x;\theta) K_{f_X,\infty}(x, X) = v(X;\theta) K_{f_X,\infty}(x, X).$$

We now show that

$$\mathbb{IF}_{2,2,\psi}(\theta) \equiv \mathbb{V}[IF_{2,2,\psi,i_1,i_2}(\theta)] = \Pi_{\theta,2} \left[ \mathbb{V} \left[ if_{1,if_{1,\psi}(O_{i_1};\cdot)}(O_{i_2};\theta)/2 \right] | \mathcal{U}_1^{\perp,2,\theta}(\theta) \right]$$

would formally have U-statistic kernel

$$IF_{2,2,\psi,i_1,i_2}(\theta) = - \begin{bmatrix} \varepsilon_{b,i_1}(\theta) E_{\theta}[H_1|X_{i_1}]^{-\frac{1}{2}} K_{f_{X,\infty}}(X_{i_1}, X_{i_2}) \\ E_{\theta}[H_1|X_{i_2}]^{-\frac{1}{2}} \varepsilon_{p,i_2}(\theta) \end{bmatrix}, \quad (15)$$

$$\text{with } \varepsilon_{b,i_1}(\theta) = \{B_{i_1}H_{1,i_1} + H_{3,i_1}\}, \quad \varepsilon_{p,i_2}(\theta) = \{H_{1,i_2}P_{i_2} + H_{2,i_2}\}.$$

To show Eq 15 note, by

$$\partial H(b, p) / \partial P = \partial \{BPH_1 + BH_2 + PH_3 + H_4\} / \partial P = BH_1 + H_3$$

and

$$\partial H(b, p) / \partial B = PH_1 + H_2,$$

we have

$$if_{1,if_{1,\psi}(O_{i_1};\cdot)}(O_{i_2};\theta) = Q_{2,b,\bar{i}_2}(\theta) + Q_{2,p,\bar{i}_2}(\theta) - IF_{1,\psi,i_2}(\theta)$$

where

$$\begin{aligned}
Q_{2,p,\bar{i}_2}(\theta) &\equiv \{B_{i_1}H_{1,i_1} + H_{3,i_1}\} if_{1,p(X_{i_1})}(O_{i_2};\theta) \\
&= -\{B_{i_1}H_{1,i_1} + H_{3,i_1}\} E_\theta[H_1|X_{i_1}]^{-\frac{1}{2}} \\
&\times K_{f_X,\infty}(X_{i_1}, X_{i_2}) E_\theta[H_1|X_{i_2}]^{-\frac{1}{2}} \{P_{i_2}H_{1,i_2} + H_{2,i_2}\} \\
&= -\varepsilon_{b,i_1}(\theta) E_\theta[H_1|X_{i_1}]^{-\frac{1}{2}} K_{f_X,\infty}(X_{i_1}, X_{i_2}) E_\theta[H_1|X_{i_2}]^{-\frac{1}{2}} \varepsilon_{p,i_2}(\theta) \\
Q_{2,b,\bar{i}_2}(\theta) &\equiv \{P_{i_1}H_{1,i_1} + H_{2,i_1}\} if_{1,b(X_{i_1})}(O_{i_2};\theta) \\
&= -\varepsilon_{b,i_2}(\theta) E_\theta[H_1|X_{i_2}]^{-\frac{1}{2}} K_{f_X,\infty}(X_{i_2}, X_{i_1}) E_\theta[H_1|X_{i_1}]^{-\frac{1}{2}} \varepsilon_{p,i_1}(\theta)
\end{aligned}$$

Thus, by part 5c of Theorem 3

$$\begin{aligned}
\mathbb{IF}_{2,2,\psi}(\theta) &= \Pi_{\theta,2} \left[ \frac{1}{2} \{Q_{2,p,\bar{i}_2}(\theta) + Q_{2,b,\bar{i}_2}(\theta) + \mathbb{IF}_{1,\psi,i_2}\} |\mathcal{U}_1^{\perp_{2,\theta}}(\theta) \right] \\
&= \frac{1}{2} \{Q_{2,p,\bar{i}_2}(\theta) + Q_{2,b,\bar{i}_2}(\theta)\} = Q_{2,p,\bar{i}_2}(\theta) \equiv \mathbb{V}[RHS \text{ of Eq. (15)}]
\end{aligned}$$

since  $\mathbb{IF}_{1,\psi,i_2}$  is a function of only one subject's data and  $Q_{2,p,\bar{i}_2}(\theta)$  and  $Q_{2,b,\bar{i}_2}(\theta)$  are the same up to a permutation that exchanges  $i_2$  with  $i_1$ .

To obtain  $IF_{3,3,\psi,\bar{i}_m}(\theta)$ , one must derive the influence function  $if_{1,if_{2,2,\psi}}(O_{i_1}, O_{i_2}; \cdot)(O_{i_3}; \theta)$  of  $if_{2,2,\psi}(O_{i_1}, O_{i_2}; \theta)$ . The formula for  $IF_{3,3,\psi,\bar{i}_m}(\theta)$  is given in Eq. (20). A detailed derivation is given in the Appendix. Here we simply note that the only essentially new point is that we now require the influence function of  $K_{f_X,\infty}(X_{i_1}, X_{i_2})$ , which, as shown next, is given by

$$IF_{1,K_{f_X,\infty}(X_{i_1}, X_{i_2})} = - \left\{ \begin{array}{c} K_{f_X,\infty}(X_{i_1}, X_{i_3}) K_{f_X,\infty}(X_{i_3}, X_{i_2}) \\ -K_{f_X,\infty}(X_{i_1}, X_{i_2}) \end{array} \right\} \quad (16)$$



To see that if Eq.(13) held, Eq.(16) would hold, note that for any path  $\tilde{\theta}(t)$  with  $\tilde{\theta}(0) = f_X(\cdot)$ ,  $h(x) = E_{\tilde{\theta}(t)} \left[ K_{\tilde{\theta}(t), \infty}(x, X_{i_1}) h(X_{i_1}) \right]$ . Differentiating wrt to  $t$  and evaluating at  $t = 0$ , we have

$$0 = E_{\theta} [K_{f_X, \infty}(x, X) h(X) S(\theta)] + E_{\theta} \left[ \left\{ \frac{\partial}{\partial t} K_{\tilde{\theta}(t), \infty}(x, X_{i_1}) \Big|_{t=0} \right\} h(X_{i_1}) \right]$$

Hence it suffices to show that

$$\begin{aligned} & - E_{\theta} [K_{f_X, \infty}(x, X) h(X) S(\theta)] \\ & = E_{\theta} [\{ E_{\theta} \{ -K_{f_X, \infty}(x, X_{i_2}) K_{f_X, \infty}(X_{i_2}, X_{i_1}) S_{i_2}(\theta) | X_{i_1} \} \} h(X_{i_1})] \end{aligned}$$

But, by Eq.(13),

$$\begin{aligned} & E_{\theta} [\{ E_{\theta} \{ -K_{f_X, \infty}(x, X_{i_2}) K_{f_X, \infty}(X_{i_2}, X_{i_1}) S_{i_2}(\theta) | X_{i_1} \} \} h(X_{i_1})] \\ & = E_{\theta} [-K_{f_X, \infty}(x, X_{i_1}) S_{i_1}(\theta) h(X_{i_1})]. \end{aligned}$$

**Feasible Estimators:** These "formal" calculations motivate a "truncated Dirac" approach to estimate  $\psi(\theta)$ . Let  $\{z_l(\cdot)\} \equiv \{z_l(X); 1, 2, \dots\}$  be a countable sequence of known basis functions with dense span in  $L_2(F_X)$  and define  $\bar{z}_k(X)^T = (z_1(X), \dots, z_k(X))$ . Define

$$K_{f_X, k}(X_{i_1}, X_{i_2}) \equiv \bar{z}_k(X_{i_1})^T \left\{ E_{f_X} \left[ \bar{z}_k(X) \bar{z}_k(X)^T \right] \right\}^{-1} \bar{z}_k(X_{i_2})$$

to be the projection kernel in  $L_2(F_X)$  onto the subspace

$$\text{lin} \{ \bar{z}_k(X) \} \equiv \{ \eta^T \bar{z}_k(x); \eta \in R^k, \eta^T \bar{z}_k(x) \in L_2(F_X) \}$$

spanned by the elements of  $\bar{z}_k(X)$ . That is, for any  $h(x)$ ,

$$\begin{aligned} & \Pi_{f_X} [h(X) | \text{lin} \{\bar{z}_k(x)\}] \\ &= E_{f_X} [K_{f_X,k}(x, X) h(X)] \\ &= \bar{z}_k(x)^T \left\{ E_{f_X} [\bar{z}_k(X) \bar{z}_k(X)^T] \right\}^{-1} E_{f_X} [\bar{z}_k(X) h(X)] \end{aligned}$$

Then we can view  $K_{f_X,k}(x_{i_1}, x_{i_2})$  as a truncated at  $k$  approximation to  $K_{f_X,\infty}(x_{i_1}, x_{i_2})$  that is in  $L_2[F_X] \times L_2[F_X]$  and satisfies Eq.(13) for all  $r(x) \in \text{lin} \{\bar{z}_k(X)\}$ . Then a natural idea would be to substitute

$$IF_{2,2,\psi,i_1,i_2}^{(k)}(\hat{\theta}) \equiv \begin{pmatrix} -\varepsilon_{b,i_1}(\hat{\theta}) E_{\hat{\theta}}[H_1|X_{i_1}]^{-\frac{1}{2}} K_{\hat{f}_X,k}(X_{i_1}, X_{i_2}) \\ \times E_{\hat{\theta}}[H_1|X_{i_2}]^{-\frac{1}{2}} \varepsilon_{p,i_2}(\hat{\theta}) \end{pmatrix}$$

with, for example,

$$\varepsilon_{b,i_1}(\hat{\theta}) = \left\{ \hat{B}_{i_1} H_{1,i_1} + H_{3,i_1} \right\}$$

for the generalized function  $IF_{2,2,\psi,i_1,i_2}(\hat{\theta})$  based on Eqs. 15 resulting in the feasible 2nd U-statistic estimator

$$\hat{\psi}_2^{(k)} = \psi(\hat{\theta}) + \mathbb{IF}_{1,\psi}(\hat{\theta}) + \mathbb{IF}_{2,2,\psi(\theta)}^{(k)}(\hat{\theta})$$

where

$$\mathbb{IF}_{2,2,\psi}^{(k)}(\hat{\theta}) \equiv \mathbb{V} \left[ IF_{2,2,\psi,i_1,i_2}^{(k)}(\hat{\theta}) \right]$$

To avoid having to do a matrix inversion it is convenient to choose  $\bar{z}_k(X) = \bar{\varphi}_k(X) / \left\{ \hat{f}_X(X) \right\}^{1/2}$  where  $\varphi_1(X), \varphi_2(X), \dots$  is a complete orthonormal basis wrt

to Lebesgue measure in  $R^d$ . Then  $E_{\hat{f}_X} [\bar{z}_k(X) \bar{z}_k(X)^T] = I_{k \times k}$  so

$$K_{\hat{f}_X, k}(X_{i_1}, X_{i_2}) = \bar{z}_k(X_{i_1})^T \bar{z}_k(X_{i_2}) = \frac{K_{Leb, k}(X_{i_1}, X_{i_2})}{\{\hat{f}_X(X_{i_1}) \hat{f}_X(X_{i_2})\}^{1/2}},$$

where  $K_{Leb, k}(X_{i_1}, X_{i_2}) \equiv \bar{\varphi}_k(X_{i_1})^T \bar{\varphi}_k(X_{i_2})$ .

This choice corresponds to having taken

$$K_{f_X, \infty}(X_{i_1}, X_{i_2}) = K_{Leb, \infty}(X_{i_1}, X_{i_2}) / \{f_X(X_{i_1}) f_X(X_{i_2})\}^{1/2}$$

in our formal calculations where  $K_{Leb, \infty}(X_{i_1}, X_{i_2})$  is the Dirac delta function wrt to Lebesgue measure. In that case with  $g(X) \equiv f_X(X) E_\theta[H_1|X]$  and  $\hat{g}(X) \equiv \hat{f}_X(X) E_{\hat{\theta}}[H_1|X]$ ,

$$IF_{2,2,\psi,i_1,i_2}(\theta) = -\varepsilon_{b,i_1}(\theta) g(X_{i_1})^{-\frac{1}{2}} K_{Leb, \infty}(X_{i_1}, X_{i_2}) g(X_{i_2})^{-\frac{1}{2}} \varepsilon_{p,i_2}(\theta) \quad (17)$$

$$IF_{2,2,\psi,i_1,i_2}^{(k)}(\hat{\theta}) = -\varepsilon_{b,i_1}(\hat{\theta}) \hat{g}(X_{i_1})^{-\frac{1}{2}} K_{Leb, k}(X_{i_1}, X_{i_2}) \hat{g}(X_{i_2})^{-\frac{1}{2}} \varepsilon_{p,i_2}(\hat{\theta}) \quad (18)$$

In the appendix, we show one can proceed by induction to formally obtain that for  $m = 3, 4, \dots$ ,

$$IF_{m,m,\psi,\bar{i}_m}(\theta) \quad (19)$$

$$= \varepsilon_{b,i_1}(\theta) g(X_{i_1})^{-\frac{1}{2}} \left[ \begin{array}{c} \sum_{j=0}^{m-2} c(m, j) \times \\ \prod_{s=1}^j \frac{H_{1,i_{s+1}}}{g(X_{i_{s+1}})} K_{Leb, \infty}(X_{i_s}, X_{i_{s+1}}) \\ \times K_{Leb, \infty}(X_{i_{j+1}}, X_{i_m}) \end{array} \right] g(X_{i_m})^{-\frac{1}{2}} \varepsilon_{p,i_m}(\theta) \quad (20)$$

where  $c(m, j) = \binom{m-2}{j} (-1)^{(j+1)}$ , which we then use to obtain  $IF_{m, m, \psi, \tilde{t}_m}^{(k)}(\hat{\theta})$  and  $\hat{\psi}_m^{(k)} = \hat{\psi}_2^{(k)} + \sum_{j=3}^m \mathbb{IF}_{j, j, \psi}^{(k)}(\hat{\theta})$ .

**Statistical Properties:** We shall prove below that the estimator  $\hat{\psi}_m^{(k)}$  has variance

$$\text{var}_{\theta} \left[ \hat{\psi}_m^{(k)} \right] \asymp \left( \frac{1}{n} \max \left[ 1, \left( \frac{k}{n} \right)^{m-1} \right] \right)$$

when  $\{\varphi_l(X); l = 1, 2, \dots\}$  is a compact wavelet basis. (Robins et al. (2007) proves this result for more general bases). We also prove that the bias

$$E_{\theta} \left[ \hat{\psi}_m^{(k)} \right] - \psi(\theta) = TB_k(\theta) + EB_m(\theta),$$

of  $\hat{\psi}_m^{(k)}$  is the sum of a truncation bias term of order

$$TB_k(\theta) = O_p(k^{-(\beta_b + \beta_p)/d})$$

(for a basis  $\{\varphi_l(X); l = 1, 2, \dots\}$  that provides optimal rate approximation for Hölder balls) and an estimation bias term of order

$$\begin{aligned} EB_m(\theta) &= O_p \left( \left\{ P - \hat{P} \right\} \left\{ B - \hat{B} \right\} \left( \frac{G - \hat{G}}{\hat{G}} \right)^{m-1} \right) \\ &= O_p \left( n^{-\frac{(m-1)\beta_g}{2\beta_g+d} - \frac{\beta_b}{2\beta_b+d} - \frac{\beta_p}{2\beta_p+d}} \right). \end{aligned}$$

The truncation bias is of this order only if  $g$  has smoothness exceeding  $\max\{\beta_p, \beta_b\}$ .

This restriction on  $g$  is removed later by using kernels based on Eq. (31). Note this estimation bias is  $O_P \left( \left\| \theta - \hat{\theta} \right\|^{m+1} \right)$ . It gets its name from the fact that, unlike

the truncation bias, it would be exactly zero if the initial estimator  $\widehat{\theta}$  happened to equal  $\theta$ . Thus, the U-statistic estimator  $\widehat{\psi}_m^{(k)}$  for our functional  $\psi(\theta)$  (which does not admit a second order influence function) differs from the U-statistic estimators  $\widehat{\psi}_m$  of Eq. (6) for functionals that admit second order influence functions in that, owing to truncation bias, the total bias of  $\widehat{\psi}_m^{(k)}$  is not  $O_p\left(\left\|\theta - \widehat{\theta}\right\|^{m+1}\right)$ . The choice of  $k$  determines the trade-off between the variance and truncation bias. As  $k \rightarrow \infty$  with  $n$  fixed,  $\text{var}_\theta\left[\widehat{\psi}_m^{(k)}\right] \rightarrow \infty$  and  $TB_k(\theta) \rightarrow 0$ . Thus, we can heuristically view the non-existent estimator  $\widehat{\psi}_m = \widehat{\psi}_m^{(k=\infty)}$  as the choice of  $k$  that results in no truncation bias [and therefore a total bias of  $O_p\left(\left\|\theta - \widehat{\theta}\right\|^{m+1}\right)$ ] at the expense of an infinite variance. Writing  $k = k(n) = n^\rho$ , the order of the asymptotic MSE of  $\widehat{\psi}_m^{(k)}$  is minimized at the value of  $\rho$  for which order of the variance equals the order of the sum of the truncation and estimation bias.

**Remark 13** *The models of examples 1-4 exhibit a spectrum of different likelihood functions and therefore a spectrum of different first order and higher order scores. Nonetheless, because the first order influence functions of the functionals  $\psi(\theta)$  share a common structure, we were able to use part 5c of Theorem 3 to formally derive  $IF_{m,m,\psi,\bar{i}_m}(\theta)$  and, thus, the feasible  $IF_{m,m,\psi,\bar{i}_m}^{(k)}(\widehat{\theta})$  in examples 1-4 in a unified manner without needing to consult the full likelihood function for any of the models. See Remark (5) above for a closely related discussion.*

**A Critical Non-uniqueness:** We have as yet neglected a critical non-uniqueness

in our definition of  $\mathbb{IF}_{m,m,\psi(\theta)}^{(k)}(\hat{\theta})$  and thus  $\hat{\psi}_m^{(k)}$  that poses a significant problem for our "truncated Dirac" approach. For instance, when  $m = 3$ , the two generalized  $U$ -statistic kernels  $IF_{3,3,\psi,i_1,i_2,i_3}(\theta)$  of Eq 20 and

$$\begin{aligned} & IF_{3,3,\psi,i_1,i_2,i_3}^*(\theta) \\ & \equiv \frac{\varepsilon_{b,i_1}(\theta)}{g(X_{i_1})^{\frac{1}{2}}} \left[ \begin{aligned} & \frac{H_{1,i_2}}{g(X_{i_2})} K_{Leb,\infty}(X_{i_1}, X_{i_2}) \\ & - E_{\theta} \left[ \frac{K_{Leb,\infty}(X_{i_1}, X_{i_2})}{f(X_{i_2})} | X_{i_1} \right] \end{aligned} \right] \\ & \times K_{Leb,\infty}(X_{i_1}, X_{i_3}) \frac{\varepsilon_{p,i_3}(\theta)}{g(X_{i_3})^{\frac{1}{2}}} \end{aligned}$$

are precisely equal, by Eq. (14); nonetheless, upon truncation, they result in different feasible kernels;

$$\begin{aligned} & IF_{3,3,\psi,i_1,i_2}^{(k)}(\hat{\theta}) \\ & = \frac{\hat{\varepsilon}_{b,i_1}(\theta)}{\hat{g}(X_{i_1})^{\frac{1}{2}}} \left[ \begin{aligned} & \frac{H_{1,i_2}}{\hat{g}(X_{i_2})} K_{Leb,k}(X_{i_1}, X_{i_2}) K_{Leb,k}(X_{i_2}, X_{i_3}) \\ & - K_{Leb,k}(X_{i_1}, X_{i_3}) \end{aligned} \right] \times \frac{\hat{\varepsilon}_{p,i_3}(\theta)}{\hat{g}(X_{i_3})^{\frac{1}{2}}} \end{aligned}$$

and

$$\begin{aligned} & IF_{3,3,\psi,i_1,i_2,i_3}^{(k),*}(\hat{\theta}) \\ & \equiv \frac{\hat{\varepsilon}_{b,i_1}(\theta)}{\hat{g}(X_{i_1})^{\frac{1}{2}}} \left[ \begin{aligned} & \frac{H_{1,i_2}}{\hat{g}(X_{i_2})} K_{Leb,k}(X_{i_1}, X_{i_2}) \\ & - E_{\hat{\theta}} \left[ \frac{K_{Leb,k}(X_{i_1}, X_{i_2})}{\hat{f}(X_{i_2})} | X_{i_1} \right] \end{aligned} \right] \\ & \times K_{Leb,k}(X_{i_1}, X_{i_3}) \frac{\hat{\varepsilon}_{p,i_3}(\theta)}{\hat{g}(X_{i_3})^{\frac{1}{2}}} \end{aligned}$$

with different orders of bias. For simplicity, we consider the case where  $H_1 = 1$  as in Examples **1a-1c**. Let  $\delta B \equiv B - \widehat{B}$ ,  $\delta P \equiv P - \widehat{P}$ ,  $\delta f = \delta g \equiv \frac{f}{\widehat{f}} - 1$ , and  $\overline{Z}_k \equiv \frac{\overline{\varphi}_k(X)}{\widehat{f}(X)^{\frac{1}{2}}} = \frac{\overline{\varphi}_k(X)}{\widehat{g}(X)^{\frac{1}{2}}}$ , then,

$$\begin{aligned}
& E_{\theta} \left[ IF_{3,3,\psi,i_1,i_2,i_3}^{(k),*} \left( \widehat{\theta} \right) \right] \\
&= E_{\theta} \left[ \begin{aligned} & \frac{\delta B_{i_1}}{\widehat{f}(X_{i_1})^{\frac{1}{2}}} \times \\ & E_{\mu} \left[ \left( \frac{f(X_{i_2})}{\widehat{f}(X_{i_2})} - 1 \right) \overline{\varphi}_k(X_{i_2})^T \right] \overline{\varphi}_k(X_{i_1}) \\ & \times E_{\theta} \left[ \frac{\delta P_{i_3}}{\widehat{f}(X_{i_3})^{\frac{1}{2}}} \overline{\varphi}_k(X_{i_3})^T \right] \overline{\varphi}_k(X_{i_1}) \end{aligned} \right] \\
&= E_{\widehat{\theta}} \left[ \begin{aligned} & \left( \frac{f(X_{i_1})}{\widehat{f}(X_{i_1})} - 1 + 1 \right) \widehat{f}(X_{i_1})^{\frac{1}{2}} \delta B_{i_1} \times \\ & E_{\widehat{\theta}} \left[ \left( \frac{f(X_{i_2})}{\widehat{f}(X_{i_2})} - 1 \right) \widehat{f}(X_{i_2})^{-\frac{1}{2}} \frac{\overline{\varphi}_k(X_{i_2})^T}{\widehat{f}(X_{i_2})^{\frac{1}{2}}} \right] \frac{\overline{\varphi}_k(X_{i_1})}{\widehat{f}(X_{i_1})^{\frac{1}{2}}} \\ & \times E_{\widehat{\theta}} \left[ \left( \frac{f(X_{i_3})}{\widehat{f}(X_{i_3})} - 1 + 1 \right) \delta P_{i_3} \frac{\overline{\varphi}_k(X_{i_3})^T}{\widehat{f}(X_{i_3})^{\frac{1}{2}}} \right] \frac{\overline{\varphi}_k(X_{i_1})}{\widehat{f}(X_{i_1})^{\frac{1}{2}}} \end{aligned} \right] \\
&= E_{\widehat{\theta}} \left[ \begin{aligned} & \widehat{f}(X_{i_1})^{\frac{1}{2}} \delta B_{i_1} E_{\widehat{\theta}} \left[ \delta f(X_{i_2}) \widehat{f}(X_{i_2})^{-\frac{1}{2}} \overline{Z}_{k,i_2}^T \right] \overline{Z}_{k,i_1} \\ & \times E_{\widehat{\theta}} \left[ \delta P_{i_3} \overline{Z}_{k,i_3}^T \right] \overline{Z}_{k,i_1} \end{aligned} \right] \\
&+ O_p \left( \left\{ B - \widehat{B} \right\} \left\{ P - \widehat{P} \right\} \left\{ G - \widehat{G} \right\}^2 \right)
\end{aligned}$$

and

$$\begin{aligned}
& E_{\theta} \left[ IF_{3,3,\psi,i_1,i_2,i_3}^{(k)} \left( \hat{\theta} \right) \right] \\
&= E_{\mu} \left[ \begin{aligned} & E_{\theta} \left[ \frac{\delta B_{i_1}}{\hat{f}(X_{i_1})^{\frac{1}{2}}} \bar{\varphi}_k(X_{i_1})^T \right] \bar{\varphi}_k(X_{i_2}) \left( \frac{f(X_{i_2})}{\hat{f}(X_{i_2})} - 1 \right) \\ & \times \bar{\varphi}_k(X_{i_2})^T E_{\theta} \left[ \frac{\delta P_{i_3}}{\hat{f}(X_{i_3})^{\frac{1}{2}}} \bar{\varphi}_k(X_{i_3}) \right] \end{aligned} \right] \\
&= E_{\hat{\theta}} \left[ \begin{aligned} & E_{\hat{\theta}} \left[ (\delta f(X_{i_1}) + 1) \delta B_{i_1} \bar{Z}_{k,i_1}^T \right] \\ & \times \bar{Z}_{k,i_2} \left( \frac{f(X_{i_2})}{\hat{f}(X_{i_2})} - 1 \right) \\ & \times \bar{Z}_{k,i_2}^T E_{\hat{\theta}} \left[ (\delta f(X_{i_3}) + 1) \delta P_{i_3} \bar{Z}_{k,i_3} \right] \end{aligned} \right] \\
&= E_{\hat{\theta}} \left[ \begin{aligned} & E_{\hat{\theta}} \left[ \delta B_{i_1} \bar{Z}_{k,i_1}^T \right] \bar{Z}_{k,i_2} \left( \frac{f(X_{i_2})}{\hat{f}(X_{i_2})} - 1 \right) \\ & \times \bar{Z}_{k,i_2}^T E_{\hat{\theta}} \left[ \delta P_{i_3} \bar{Z}_{k,i_3} \right] \end{aligned} \right] \\
&+ O_p \left( \left\{ B - \hat{B} \right\} \left\{ P - \hat{P} \right\} \left\{ G - \hat{G} \right\}^2 \right)
\end{aligned}$$

That is,

$$\begin{aligned}
& E_{\theta} \left[ IF_{3,3,\psi,i_1,i_2,i_3}^{(k),*} \left( \hat{\theta} \right) \right] - E_{\theta} \left[ IF_{3,3,\psi,i_1,i_2,i_3}^{(k)} \left( \hat{\theta} \right) \right] \\
&= E_{\hat{\theta}} \left[ \begin{aligned} & \hat{f}(X_{i_1})^{\frac{1}{2}} \delta B_{i_1} E_{\hat{\theta}} \left[ \delta f(X_{i_2}) \hat{f}(X_{i_2})^{-\frac{1}{2}} \bar{Z}_{k,i_2}^T \right] \bar{Z}_{k,i_1} \\ & E_{\hat{\theta}} \left[ \delta P_{i_3} \bar{Z}_{k,i_3}^T \right] \bar{Z}_{k,i_1} \end{aligned} \right] \\
&- E_{\hat{\theta}} \left[ \begin{aligned} & E_{\hat{\theta}} \left[ \delta B_{i_1} \bar{Z}_{k,i_1}^T \right] \bar{Z}_{k,i_2} \delta f(X_{i_2}) \\ & \times \bar{Z}_{k,i_2}^T E_{\hat{\theta}} \left[ \delta P_{i_3} \bar{Z}_{k,i_3} \right] \end{aligned} \right] \\
&+ O_p \left( \left\{ B - \hat{B} \right\} \left\{ P - \hat{P} \right\} \left\{ G - \hat{G} \right\}^2 \right)
\end{aligned}$$

=



$$\begin{aligned}
& E_{\hat{\theta}} \left[ \Pi_{\hat{\theta}} [\delta P | \bar{Z}_k] \hat{f}(X)^{\frac{1}{2}} \delta B \Pi_{\hat{\theta}} \left[ \delta f(X) \hat{f}(X)^{-\frac{1}{2}} | \bar{Z}_k \right] \right] \\
& - E_{\hat{\theta}} \left[ \Pi_{\hat{\theta}} [\delta P | \bar{Z}_k] \hat{f}(X)^{\frac{1}{2}} \delta f(X) \hat{f}(X)^{-\frac{1}{2}} \Pi_{\hat{\theta}} [\delta B | \bar{Z}_k] \right] \\
& + O_p \left( \left\{ B - \hat{B} \right\} \left\{ P - \hat{P} \right\} \left\{ G - \hat{G} \right\}^2 \right) \\
& = \\
& E_{\hat{\theta}} \left[ \begin{aligned} & \Pi_{\hat{\theta}} [\delta P | \bar{Z}_k] \hat{f}(X)^{\frac{1}{2}} \times \\ & \left\{ \begin{aligned} & \Pi_{\hat{\theta}}^{\perp} [\delta B | \bar{Z}_k] \Pi_{\hat{\theta}} \left[ \delta f(X) \hat{f}(X)^{-\frac{1}{2}} | \bar{Z}_k \right] \\ & - \Pi_{\hat{\theta}} [\delta B | \bar{Z}_k] \Pi_{\hat{\theta}}^{\perp} \left[ \delta f(X) \hat{f}(X)^{-\frac{1}{2}} | \bar{Z}_k \right] \end{aligned} \right\} \end{aligned} \right] \\
& + O_p \left( \left\{ B - \hat{B} \right\} \left\{ P - \hat{P} \right\} \left\{ G - \hat{G} \right\}^2 \right)
\end{aligned}$$

where  $\Pi_{\hat{\theta}} [h(X) | \bar{Z}_k]$  and  $\Pi_{\hat{\theta}}^{\perp} [h(X) | \bar{Z}_k]$  respectively denote the projection under  $F(\cdot; \hat{\theta})$  in  $L_2(\hat{F})$  of  $h(X)$  on the  $k$  dimensional linear subspace  $\text{lin} \{ \bar{z}_k(X) \}$  spanned by the components of the vector  $\bar{z}_k(X)$  and the projection on the orthocomplement of this subspace.

Since the basis  $\{ \varphi_l(X) ; l = 1, 2, \dots \}$  provides optimal rate approximation for Hölder balls, it is easy to verify that the difference is of order

$$O_p \left( \begin{aligned} & n^{-\frac{\beta_p/d}{1+2\beta_p/d} - \frac{\beta_g/d}{1+2\beta_g/d}} k^{-\beta_b/d} + n^{-\frac{\beta_p/d}{1+2\beta_p/d} - \frac{\beta_b/d}{1+2\beta_b/d}} k^{-\beta_g/d} \\ & + n^{-\frac{\beta_p/d}{1+2\beta_p/d} - \frac{\beta_b/d}{1+2\beta_b/d} - \frac{2\beta_g/d}{1+2\beta_g/d}} \end{aligned} \right)$$

provided  $g$  has smoothness exceeding  $\max(\beta_p, \beta_b)$ .

For concreteness, we shall look at an example. Suppose  $\beta_b/d = \beta_p/d = 0.3$  and  $\beta_g/d = 0.1$ , thus, by choosing  $k = n^{\frac{5}{6}}, \hat{\psi}_3^{(k)}$  converges to  $\psi(\theta)$  at rate  $n^{-\frac{1}{2}}$ . In contrast,

the order,

$$\min_k \left( n^{-\frac{\beta_p/d}{1+2\beta_p/d} - \frac{\beta_g/d}{1+2\beta_g/d}} k^{-\beta_b/d} + n^{-\frac{\beta_p/d}{1+2\beta_p/d} - \frac{\beta_b/d}{1+2\beta_b/d}} k^{-\beta_g/d} + \sqrt{\frac{1}{n} \max \left( 1, \frac{k^2}{n^2} \right)} \right),$$

of the optimal root mean squares error of  $\widehat{\psi}_3^{(k),*}$  that uses  $IF_{3,3,\psi,\bar{i}_3}^{(k),*}(\widehat{\theta})$  is  $n^{-0.477} \gg n^{-0.5}$ . Thus  $\widehat{\psi}_3^{(k),*}$  converges to  $\psi(\theta)$  at a slower rate than  $\widehat{\psi}_3^{(k)}$  which uses  $IF_{3,3,\psi,\bar{i}_3}^{(k)}(\widehat{\theta})$ .

Nothing in our development up to this point provides any guidance as to which of the many equivalent generalized U-statistic kernels should be selected for truncation.

To provide some guidance, we introduce an alternative approach to the estimation of  $\psi(\theta)$  based on truncated parameters that admit higher order influence functions. The class of estimators we derive using this alternative approach includes members algebraically identical to the estimators  $\widehat{\psi}_m^{(k)}$  but does not include estimators equivalent to less efficient estimators such as  $\widehat{\psi}_3^{(k),*}$ .

**An Approach based on Truncated Parameters:** We introduce a class of truncated parameters  $\widetilde{\psi}_k(\theta)$  that (i) depend on the sample size through a positive integer index  $k = k(n)$  (which we refer to as the truncation index and will be optimized below), (ii) have influence functions  $\mathbb{IF}_{m,\widetilde{\psi}_k}(\theta)$  of all orders  $m$ , (iii) equals  $\psi(\theta)$  on a large subset  $\Theta_{sub,k}$  of  $\Theta$  and (iv) the initial estimator  $\widehat{\theta}$  is an element of  $\Theta_{sub,k}$  so that the plug-ins  $\psi(\widehat{\theta})$  and  $\widetilde{\psi}_k(\widehat{\theta})$  are equal. To prepare we introduce a simplified notation. For functions  $h(o, \cdot)$  or  $r(\cdot)$  of  $\theta$ , we will often write  $h(o, \widehat{\theta})$  and  $r(\widehat{\theta})$  as  $\widehat{h}(o)$  and  $\widehat{r}$ , and  $E_{\widehat{\theta}}[\cdot]$  as  $\widehat{E}[\cdot]$ . Similarly, we often write  $h(o, \theta)$  and  $r(\theta)$  as  $h(o)$  and  $r$ , and  $E_{\theta}[\cdot]$  as  $E[\cdot]$ . Further we shall introduce slightly different definitions

of truncation and estimation bias.

Define the estimator  $\psi_{m,k}(\hat{\theta}) \equiv \psi(\hat{\theta}) + \mathbb{IF}_{m,\tilde{\psi}_k}(\hat{\theta})$  or, equivalently,  $\hat{\psi}_{m,k} \equiv \hat{\psi} + \hat{\mathbb{IF}}_{m,\tilde{\psi}_k}$ . Then the conditional bias  $E[\hat{\psi}_{m,k}|\hat{\theta}] - \psi$  of  $\hat{\psi}_{m,k}$  is  $TB_k + EB_{m,k}$ , where the truncation bias  $TB_k = \tilde{\psi}_k - \psi$  is zero for  $\theta \in \Theta_{sub,k}$  and does not depend on  $m$  and the estimation bias  $EB_{m,k} = E[\hat{\psi}_{m,k}|\hat{\theta}] - \tilde{\psi}_k$  is  $O_P(\|\hat{\theta} - \theta\|^{m+1})$  by Theorem 2. Since, as we show later, the order of  $EB_{m,k}$  does not depend on  $k$ , we will abbreviate  $EB_{m,k}$  as  $EB_m$ , suppressing the dependence on  $k$ . Under minimal conditions, the conditional variance of  $\hat{\psi}_{m,k}$  is of the order of  $\text{var}[\mathbb{IF}_{m,\tilde{\psi}_k}]$  whenever  $k \equiv k(n) \geq n$ . The rate of convergence of  $\hat{\psi}_{m,k}$  to  $\psi$  can depend on the choice of  $\tilde{\psi}_k$ . Nevertheless, many different choices  $\tilde{\psi}_k$  result in estimators  $\hat{\psi}_{m,k}$  that achieve what we conjecture to be the optimal rate for estimators in our class of the form  $\hat{\psi}_{m,k}$ . We choose, among all such  $\tilde{\psi}_k$ , the class that minimizes the computational complexity of  $\hat{\psi}_{m,k}$ . Specifically for all  $\tilde{\psi}_k$  in our chosen class and all  $j$ ,  $\mathbb{IF}_{jj,\tilde{\psi}_k}$  consists of a single term rather than a sum of many terms. We conjecture this appealing property does not hold for any  $\tilde{\psi}_k$  outside our class. We now describe this choice. The parameter  $\tilde{\psi}_k$  is defined in terms of  $k(n)$ -dimensional 'working' linear parametric submodels for  $p(\cdot)$  and  $b(\cdot)$  depending on unknown parameters  $\bar{\alpha}_k$  and  $\bar{\eta}_K$  through the basepoints  $\hat{p}(\cdot)$  and  $\hat{b}(\cdot)$ , where  $\hat{p}(\cdot)$  and  $\hat{b}(\cdot)$  are initial estimators from the training sample. Specifically let  $\dot{p}(X)$  and  $\dot{b}(X)$  be arbitrary bounded known functions chosen by the analyst satisfying Eqs (21) – (23) below.

$$\dot{p}(X) \dot{b}(X) E[H_1|X] \geq 0 \text{ w.p.1} \quad (21)$$

$$\left\| \frac{\dot{p}(X)}{\dot{b}(X)} \right\|_{\infty} < C^*, \left\| \frac{\dot{b}(X)}{\dot{p}(X)} \right\|_{\infty} < C^* \quad (22)$$

$$\frac{\dot{p}(X)}{\dot{b}(X)} \text{ has at least } [\max\{\beta_b, \beta_p\}] \text{ derivatives} \quad (23)$$

Particular choices of  $\dot{p}(X)$  and  $\dot{b}(X)$  can make the form of  $\mathbb{IF}_{m, \tilde{\psi}_k}(\hat{\theta})$  more aesthetic. The choice has no bearing on the rate of convergence of the estimator  $\hat{\psi}_{m,k}$  to  $\psi(\theta)$ . Often there are fairly natural choices for  $\dot{p}(\cdot)$  and  $\dot{b}(\cdot)$ . See Remark 16 below for examples. Let  $\bar{\alpha}_k, \bar{\eta}_k$  be  $k$ -vectors of unknown parameters and consider the 'working' linear models

$$p^*(X, \bar{\alpha}_k) \equiv \hat{p}(X) + \dot{p}(X) \bar{\alpha}_k^T \bar{z}_k(X) \equiv \hat{P} + \dot{P} \bar{\alpha}_k^T \bar{Z}_k \quad (24)$$

$$b^*(X, \bar{\eta}_k) = \hat{b}(X) + \dot{b}(X) \bar{\eta}_k^T \bar{z}_k(X) = \hat{B} + \dot{B} \bar{\eta}_k^T \bar{Z}_k \quad (25)$$

We define the parameters  $\tilde{\eta}_k(\theta)$  and  $\tilde{\alpha}_k(\theta)$  respectively to be the solution to

$$0 = E_{\theta} [\partial H(b^*(X, \bar{\eta}_k), p^*(X, \bar{\alpha}_k)) / \partial \bar{\alpha}_k] = E_{\theta} [\{H_1 b^*(X, \bar{\eta}_k) + H_3\} \dot{P} \bar{Z}_k] \quad (26)$$

$$0 = E_{\theta} [\partial H(b^*(X, \bar{\eta}_k), p^*(X, \bar{\alpha}_k)) / \partial \bar{\eta}_k] = E_{\theta} [\{H_1 p^*(X, \bar{\alpha}_k) + H_2\} \dot{B} \bar{Z}_k]. \quad (27)$$

The solution to (26) and (27) exist in closed form as

$$\tilde{\eta}_k(\theta) = -E_{\theta} [\dot{B} \dot{P} H_1 \bar{Z}_k \bar{Z}_k^T]^{-1} E_{\theta} [\bar{Z}_k \dot{P} \{H_1 \hat{B} + H_3\}] \quad (28)$$

$$\tilde{\alpha}_k(\theta) = -E_{\theta} [\dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T]^{-1} E_{\theta} [\bar{Z}_k \dot{B} \{H_1 \hat{P} + H_2\}]. \quad (29)$$

Next define  $\tilde{b}(\theta) = \tilde{b}(\cdot, \theta) = b^*(\cdot, \tilde{\eta}_k(\theta))$  and  $\tilde{p}(\theta) = \tilde{p}(\cdot, \theta) = p^*(\cdot, \tilde{\alpha}_k(\theta))$  and

$$\tilde{\psi}_k(\theta) = E_\theta \left[ H \left( \tilde{b}(\theta), \tilde{p}(\theta) \right) \right]$$

Note the models  $p^*(\cdot, \bar{\alpha}_k)$  and  $b^*(\cdot, \bar{\eta}_k)$  are used only to define the truncated parameter  $\tilde{\psi}_k(\theta)$ . They are not assumed to be correctly specified. In particular, the training sample estimates  $\hat{p}, \hat{b}$  need not be based on the models  $p^*(\cdot, \bar{\alpha}_k), b^*(\cdot, \bar{\eta}_k)$ . We now compare our truncated parameter  $\tilde{\psi}_k(\theta)$  with  $\psi(\theta)$  and calculate the truncation bias. It is important to keep in mind that  $b, p$  are components of the unknown  $\theta$  while  $\hat{p}, \hat{b}, \hat{p}, \hat{b}$  are regarded as known functions.

**Theorem 14** *If our model satisfies Aii) – Aiii) and*

$$\theta \in \Theta_{sub,k} = \{\theta; p(\cdot) = p^*(\cdot, \bar{\alpha}_k) \text{ for some } \bar{\alpha}_k \text{ or } b(\cdot) = b^*(\cdot, \bar{\eta}_k) \text{ for some } \bar{\eta}_k\} \cap \Theta$$

*then*  $\tilde{\psi}_k(\theta) = \psi(\theta)$

$$\text{Further } TB_k(\theta) = \tilde{\psi}_k(\theta) - \psi(\theta) = E_\theta \left[ \left\{ \tilde{B}(\theta) - B \right\} \left\{ \tilde{P}(\theta) - P \right\} H_1 \right]$$

**Proof.** *Immediate from Theorem 9 and Lemma 10. ■*

We know from the above Theorem that  $TB_k(\theta) = 0$  for  $\theta \in \Theta_{sub,k}$ . However to control the truncation bias in forming confidence intervals for  $\psi(\theta)$  we will need to know how fast  $\sup_{\theta \in \Theta} \{TB_k(\theta)\}$  decreases as  $k$  increases. The following theorem is a key step towards determining an upper bound.

**Theorem 15** Suppose  $\dot{b}(X)$  and  $\dot{p}(X)$  are chosen so that  $\dot{B}\dot{P}E[H_1|X] \geq 0$  wp1. Let

$$Q \equiv q(X) = \left\{ \dot{B}\dot{P}E[H_1|X] \right\}^{1/2}$$

and  $\Pi[h(Z)|Q\bar{Z}_k]$  and  $\Pi^\perp[h(X)|Q\bar{Z}_k]$  be, respectively, the projection in  $L_2(F_X(x))$  of  $h(X)$  on the  $k$  dimensional linear subspace  $\text{lin}\{Q\bar{Z}_k\}$  spanned by the components of the vector  $Q\bar{Z}_k = q(X)\bar{z}_k(X)$  and the projection on the orthocomplement of this subspace. Then if Ai) – Aiii) are satisfied ,

$$TB_k = E \left[ \Pi^\perp \left[ \left( \frac{P - \hat{P}}{\dot{P}} \right) Q|Q\bar{Z}_k \right] \Pi^\perp \left[ \left( \frac{B - \hat{B}}{\dot{B}} \right) Q|Q\bar{Z}_k \right] \right]$$

**Remark 16** To simplify various formulae it is often convenient and aesthetically pleasing to have  $\hat{Q} = 1$ . We can choose  $\dot{B}$  and  $\dot{P}$  to guarantee  $\hat{Q} = 1$  wp1. For the functional  $\psi(\theta) = E_\theta[b(X)p(X)]$  of Example 1a,  $H_1 = -1$  wp1. Thus choosing  $\dot{B}$  and  $\dot{P}$  equal to 1 and  $-1$ , respectively, wp1 makes  $\hat{Q} = 1$  wp1. In the missing data example 2a, the function  $H_1 = -A$  so  $\hat{E}[H_1|X] = 1/\hat{P}$  and thus the choice  $\dot{B} = -1, \dot{P} = \hat{P}$  makes  $\hat{Q} = 1$  wp1. Note since inference on  $\psi(\theta)$  is conditional on the training sample data, we view the initial estimator  $\hat{p}(\cdot)$  of  $p(\cdot)$  from the training sample as known and thus an analyst is free to choose  $\dot{P}$  to be  $\hat{P}$ .

**Examples continued.** In **Example 1a**, recall  $\psi = E[BP]$ . Choose  $\dot{B} = -\dot{P} =$

1 *wp1* so  $\widehat{Q} = Q = 1$ , and take  $\widehat{B} \in \text{lin} \{ \overline{Z}_k \}$ . Then

$$\begin{aligned}\widetilde{B} &= \widehat{B} + \Pi \left[ \left( B - \widehat{B} \right) | \overline{Z}_k \right] = \Pi \left[ B | \overline{Z}_k \right] \\ \widetilde{P} &= \Pi \left[ P | \overline{Z}_k \right], \\ TB_k &= E \left\{ \left[ \Pi^\perp \left[ B | \overline{Z}_k \right] \Pi^\perp \left[ P | \overline{Z}_k \right] \right] \right\}, \\ \widetilde{\psi}_k &= \psi - TB_k = E \left\{ \Pi \left[ B | \overline{Z}_k \right] \Pi \left[ P | \overline{Z}_k \right] \right\}\end{aligned}$$

Thus  $\widetilde{\psi}_k$  appears to be the natural choice for a truncated parameter.

In the **Example 2a** with  $\psi = E[B]$ ,  $\dot{B} = -1$ ,  $\dot{P} = \widehat{P} = 1/\widehat{\pi}$ ,  $\widehat{Q} = 1$ ,  $Q = \left\{ \widehat{P}/P \right\}^{1/2} = \left\{ \frac{\pi}{\widehat{\pi}} \right\}^{1/2}$ ,  $\widehat{\pi} \equiv \widehat{\pi}(X)$ ,  $\pi \equiv \pi(X)$ , we obtain

$$TB_k = E \left[ \begin{array}{l} \Pi^\perp \left[ \widehat{\pi} \left( \frac{1}{\pi} - \frac{1}{\widehat{\pi}} \right) \left\{ \frac{\pi}{\widehat{\pi}} \right\}^{1/2} | \left\{ \frac{\pi}{\widehat{\pi}} \right\}^{1/2} \overline{Z}_k \right] \\ \times \Pi^\perp \left[ \left\{ \frac{\pi}{\widehat{\pi}} \right\}^{1/2} \left( B - \widehat{B} \right) | \left\{ \frac{\pi}{\widehat{\pi}} \right\}^{1/2} \overline{Z}_k \right] \end{array} \right]$$

Thus the truncated parameter  $\widetilde{\psi}_k = \psi - TB_k$  does not seem to be a particular natural or obvious choice. The complexity of  $\widetilde{\psi}_k$  is not simply due to the fact that we chose  $\dot{P} = \widehat{P}$  rather than  $\dot{P} = 1$  as we now demonstrate.

In **Example 2a** with  $\dot{B} = -1$ ,  $\dot{P} = 1$ ,  $\widehat{Q} = \widehat{\pi}^{1/2}$ ,  $Q = \pi^{1/2}$ ,

$$TB_k = E \left[ \begin{array}{l} \Pi^\perp \left[ \left( \frac{1}{\pi} - \frac{1}{\widehat{\pi}} \right) \pi^{1/2} | \pi^{1/2} \overline{Z}_k \right] \times \\ \Pi^\perp \left[ \left\{ \frac{\pi}{\widehat{\pi}} \right\}^{1/2} \left( B - \widehat{B} \right) | \pi^{1/2} \overline{Z}_k \right] \end{array} \right]$$

Nonetheless we will see that, for either choice of  $(\dot{B}, \dot{P})$ , the parameter  $\widetilde{\psi}_k$  will result in estimators with good properties.

**Remark 17** Henceforth, given  $(\beta_p, \beta_b, \beta_g)$ ,  $\{\varphi_l(X), l = 1, 2, \dots\}$  will always denote a complete orthonormal basis wrt to Lebesgue measure in  $R^d$  or in the unit cube in  $R^d$  that provides optimal rate approximation for Hölder balls  $H(\beta^*, C), \beta^* \leq [\max(\beta_p, \beta_b, \beta_g)]$ , i.e.

$$\sup_{h \in H(\beta^*, C)} \inf_{\varsigma_l} \int_{R^d} \left( h(x) - \sum_{l=1}^k \varsigma_l \varphi_l(x) \right)^2 dx = O(k^{-2\beta^*/d}) \quad (30)$$

The basis consisting of  $d$ -fold tensor products of univariate orthonormal polynomials satisfies (30) for all  $\beta^*$ . The basis consisting of  $d$ -fold tensor products of a univariate Daubechies compact wavelet basis with mother wavelet  $\varphi_w(u)$  satisfying

$$\int_{R^1} u^m \varphi_w(u) du = 0, m = 0, 1, \dots, M$$

also satisfies (30) for  $\beta^* < M + 1$ .

**Theorem 18** Suppose that  $A_i) - A_{iv})$  are satisfied, that  $\dot{b}(X)$  and  $\dot{p}(X)$  satisfy (21) – (23) and in the remainder of the paper, unless stated otherwise, we take

$$\bar{z}_k(X) \equiv E \left\{ \widehat{Q}^2 \bar{\varphi}_k(X) \bar{\varphi}_k(X)^T \right\}^{-1/2} \bar{\varphi}_k(X) \quad (31)$$

where recall that  $\widehat{Q}^2 = \left\{ \dot{B} \dot{P} \widehat{E}[H_1|X] \right\}$ . Then

$$\sup_{\theta \in \Theta} \{TB_k^2(\theta)\} = O_p(k^{-2(\beta_b + \beta_p)/d})$$



### 3.2.2 Derivation of the Higher Order Influence Functions of the Truncated Parameter

We begin by proving that the first order influence functions of  $\tilde{\psi}_k$  and  $\psi$  are identical except with  $\tilde{b}(\theta), \tilde{p}(\theta), \tilde{\psi}_k(\theta)$  replacing  $b, p, \psi(\theta)$ .

#### Theorem 19

$$\mathbb{IF}_{1, \tilde{\psi}_k}(\theta) = \mathbb{V} \left[ IF_{1, \tilde{\psi}_k, i_1}(\theta) \right]$$

with

$$IF_{1, \tilde{\psi}_k}(\theta) = H \left( \tilde{b}(\theta), \tilde{p}(\theta) \right) - \tilde{\psi}_k(\theta)$$

**Proof.** Since  $\tilde{\psi}_k(\theta) = E_\theta \left[ H \left( \tilde{b}(\theta), \tilde{p}(\theta) \right) \right]$ ,

$$\begin{aligned} IF_{1, \tilde{\psi}_k}(\theta) &= H \left( \tilde{b}(\theta), \tilde{p}(\theta) \right) - \tilde{\psi}_k(\theta) \\ &+ E \left[ \partial H \left( b^* \left( X, \tilde{\eta}_k(\theta) \right), p^* \left( X, \tilde{\alpha}_k(\theta) \right) \right) / \partial \tilde{\eta}_k^T \right] IF_{1, \tilde{\eta}_k(\cdot)}(\theta) \\ &+ E \left[ \partial H \left( b^* \left( X, \tilde{\eta}_k(\theta) \right), p^* \left( X, \tilde{\alpha}_k(\theta) \right) \right) / \partial \tilde{\alpha}_k^T \right] IF_{1, \tilde{\alpha}_k(\cdot)}(\theta) \end{aligned}$$

But, by definition of  $\tilde{\eta}_k(\theta)$  and  $\tilde{\alpha}_k(\theta)$ , both expectations are zero. ■

Note that  $\tilde{\eta}_k(\theta)$  and  $\tilde{\alpha}_k(\theta)$  are not maximizers of the expected log-likelihood for  $\bar{\alpha}_k$  and  $\bar{\eta}_k$ . This choice was deliberate. Had we defined  $\tilde{\eta}_k(\theta)$  and  $\tilde{\alpha}_k(\theta)$  as the maximizers of the expected log-likelihood, then  $\mathbb{IF}_{1, \tilde{\psi}_k}(\theta)$  would have had additional terms since the expectations in the preceding proof would not be zero. The existence of these extra terms would translate to many extra terms in  $\mathbb{IF}_{m, \tilde{\psi}_k}(\theta)$  for large  $m$

leading to computational difficulties. Similarly had we chosen models  $p^*(X, \bar{\alpha}_k) \equiv \Phi(\hat{P} + \dot{P}\bar{\alpha}_k^T \bar{Z}_k)$  and  $b^*(X, \bar{\eta}_k) = \Phi(\hat{B} + \dot{B}\bar{\eta}_k^T \bar{Z}_k)$  with  $\Phi(\cdot)$  a non-linear inverse-link function,  $\mathbb{IF}_{m, \tilde{\psi}_k}(\theta)$  would also have had many extra terms without an improvement in the rate of convergence.

The following is proved in the Appendix.

**Theorem 20**  $\mathbb{IF}_{m, \tilde{\psi}_k} = \mathbb{IF}_{1, \tilde{\psi}_k} + \sum_{j=2}^m \mathbb{IF}_{jj, \tilde{\psi}_k}$  where  $\mathbb{IF}_{jj, \tilde{\psi}_k} = \mathbb{V}[IF_{jj, \tilde{\psi}_k, \bar{i}_j}]$  is a  $j$ th order degenerate  $U$ -statistic given by

$$IF_{22, \tilde{\psi}_k, \bar{i}_2} = - \left\{ \begin{aligned} & \left[ (H_1 \tilde{P} + H_2) \dot{B} \bar{Z}_k^T \right]_{i_1} \left\{ E \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} \\ & \times \left[ \bar{Z}_k (H_1 \tilde{B} + H_3) \dot{P} \right]_{i_2} \end{aligned} \right\}$$

$$IF_{jj, \tilde{\psi}_k, \bar{i}_j} = (-1)^{j-1} \left[ (H_1 \tilde{P} + H_2) \dot{B} \bar{Z}_k^T \right]_{i_1}$$

$$\times \left[ \begin{aligned} & \prod_{s=3}^j \left\{ E \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} \\ & \left\{ \left( \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right)_{i_s} - E \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\} \end{aligned} \right]$$

$$\times \left\{ E \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} \left[ \bar{Z}_k (H_1 \tilde{B} + H_3) \dot{P} \right]_{i_2}$$

### 3.2.3 The Estimator $\hat{\psi}_{m,k} \equiv \hat{\psi} + \widehat{\mathbb{IF}}_{m, \tilde{\psi}_k}$ and its Estimation Bias

We can now calculate the estimator  $\hat{\psi}_{m,k} \equiv \hat{\psi} + \widehat{\mathbb{IF}}_{m, \tilde{\psi}_k}$  by substitution of  $\hat{\theta}$  for  $\theta$  in

$\mathbb{IF}_{m, \tilde{\psi}_k} \equiv \mathbb{IF}_{m, \tilde{\psi}_k}(\theta)$  to obtain the following.

**Theorem 21** Suppose (31) holds . Then  $\widehat{\psi}_{m,k} = \widehat{\psi} + \widehat{\mathbb{I}\mathbb{F}}_{1,\widetilde{\psi}_k} + \sum_{j=2}^m \widehat{\mathbb{I}\mathbb{F}}_{jj,\widetilde{\psi}_k}$  where

$$\begin{aligned}\widehat{\psi} + \widehat{\mathbb{I}\mathbb{F}}_{1,\widetilde{\psi}_k} &= \widehat{B}\widehat{P}H_1 + \widehat{B}H_2 + \widehat{P}H_3 + H_4 \\ \widehat{IF}_{22,\widetilde{\psi}_k,\bar{i}_2} &= - \left[ \left( H_1\widehat{P} + H_2 \right) \dot{B}\overline{Z}_k^T \right]_{i_1} \left[ \overline{Z}_k \left( H_1\widehat{B} + H_3 \right) \dot{P} \right]_{i_2} \\ \widehat{IF}_{jj,\widetilde{\psi}_k,\bar{i}_j} &= (-1)^{j-1} \left\{ \left[ \left( H_1\widehat{P} + H_2 \right) \dot{B}\overline{Z}_k^T \right]_{i_1} \left[ \prod_{s=3}^j \left\{ \begin{array}{c} \left( \dot{P}\dot{B}H_1\overline{Z}_k\overline{Z}_k^T \right)_{i_s} \\ -I_{k \times k} \end{array} \right\} \right] \right\} \\ &\quad \times \left[ \overline{Z}_k \left( H_1\widehat{B} + H_3 \right) \dot{P} \right]_{i_2} \end{aligned}$$

**Proof.** By Lemma 10  $E_{\widehat{\theta}} \left[ \left\{ H_1\widehat{B} + H_3 \right\} \dot{P}\overline{Z}_k \right] = E_{\widehat{\theta}} \left[ \left\{ H_1\widehat{P} + H_2 \right\} \dot{B}\overline{Z}_k \right] = 0$ . Thus by Eqs. (28) and (29)  $\widetilde{\eta}_k(\widehat{\theta}) = \widetilde{\alpha}_k(\widehat{\theta}) = 0$  so  $\widetilde{B}(\widehat{\theta}) = \widehat{B}$  and  $\widetilde{P}(\widehat{\theta}) = \widehat{P}$ . Further, by Eq.(31),  $\widehat{E} \left[ \dot{P}\dot{B}H_1\overline{Z}_k\overline{Z}_k^T \right] = \widehat{E} \left[ \dot{P}\dot{B}\widehat{E}[H_1|X] \overline{Z}_k\overline{Z}_k^T \right] = \widehat{E} \left[ \widehat{Q}^2\overline{Z}_k\overline{Z}_k^T \right] = \int \overline{\varphi}_k(x) \overline{\varphi}_k(x)^T = I_{k \times k}$ . ■

It follows that by our judicious choice of  $\overline{Z}_k$  in Eq.(31), we have avoided the need to invert a  $k \times k$  matrix to compute  $\widehat{\psi}_{m,k}$ .

**Remark 22** The reader can easily check that when we take

$$\overline{Z}_k = \overline{\varphi}_k(X) / \left\{ \widehat{f}_X(X) \widehat{E}[H_1|X] \right\}^{1/2}$$

$\dot{B} = \dot{P} = 1$  and  $H_1 \geq 0$  w.p 1,  $\widehat{IF}_{j,j,\widetilde{\psi}_k,\bar{i}_2}$  is precisely the same as  $IF_{j,j,\psi,i_1,i_2}^{(k)}(\widehat{\theta})$  of equation (18) in Section 3.2.1.

To make our procedures less abstract, we provide explicit expressions for  $\widehat{IF}_{jj,\widetilde{\psi}_k,\bar{i}_j}$  in examples 1a and 2a.

**Example 1a continued:**  $\psi = E[BP]$ ,  $\dot{B} = -\dot{P} = 1$  wpl,  $\hat{Q} = 1$ ,  $H_1 = -1$ ,  
 $\hat{E}[\bar{Z}_k \bar{Z}_{ki_s}^T] = I_{k \times k}$ . Then

$$\dot{P} \{H_1 \hat{B} + H_3\} = Y - \hat{B}, \dot{B} \{H_1 \hat{P} + H_2\} = A - \hat{P}$$

and thus

$$\begin{aligned} \widehat{IF}_{22, \tilde{\psi}_k, \tilde{i}_2} &= - \left[ (A - \hat{P}) \bar{Z}_k^T \right]_{i_1} \left[ \bar{Z}_k (Y - \hat{B}) \right]_{i_2} \\ \widehat{IF}_{jj, \tilde{\psi}_k, \tilde{i}_j} &= (-1)^{j-1} \left[ (A - \hat{P}) \bar{Z}_k^T \right]_{i_1} \left[ \prod_{s=3}^j \left\{ \bar{Z}_{ki_s} \bar{Z}_{ki_s}^T - I_{k \times k} \right\} \right] \left[ \bar{Z}_k (Y - \hat{B}) \right]_{i_2} \end{aligned}$$

**Example 2a continued:**  $H_1 = -A$ ,  $\dot{B} = -1$ ,  $\dot{P} = \hat{P} = 1/\hat{\pi}$ ,  $\hat{Q} = 1$ ,  $\psi = E[B]$ ,  
 $Q = \{\hat{P}/P\}^{1/2} = \{\frac{\pi}{\hat{\pi}}\}^{1/2}$  and  $\bar{Z}_k = \bar{\varphi}_k \{ \hat{f}(X) \}^{-1/2}$ , so  $\hat{E}[\bar{Z}_k \bar{Z}_{ki_s}^T] = I_{k \times k}$ .

Then  $\dot{P} \{H_1 \hat{B} + H_3\} = \frac{A}{\hat{\pi}} (Y - \hat{B})$ ,  $\dot{B} \{H_1 \hat{P} + H_2\} = (\frac{A}{\hat{\pi}} - 1)$ , so

$$\begin{aligned} \widehat{IF}_{22, \tilde{\psi}_k, \tilde{i}_2} &= - \left[ \left( \frac{A}{\hat{\pi}} - 1 \right) \bar{Z}_k^T \right]_{i_1} \left[ \bar{Z}_k \frac{A}{\hat{\pi}} (Y - \hat{B}) \right]_{i_2} \\ \widehat{IF}_{jj, \tilde{\psi}_k, \tilde{i}_j} &= (-1)^{j-1} \left[ \left( \frac{A}{\hat{\pi}} - 1 \right) \bar{Z}_k^T \right]_{i_1} \left[ \prod_{s=3}^j \left\{ \frac{A}{\hat{\pi}} \bar{Z}_k \bar{Z}_k^T - I_{k \times k} \right\}_{i_s} \right] \left[ \bar{Z}_k \frac{A}{\hat{\pi}} (Y - \hat{B}) \right]_{i_2} \end{aligned}$$

Consider Example 2a with  $\dot{B} = -1$ ,  $\dot{P} = 1$ ,  $\hat{Q} = \hat{\pi}^{1/2}$ ,  $\hat{E}[\hat{Q}^2 \bar{Z}_k \bar{Z}_{ki_s}^T] = I_{k \times k}$ ,  
 $\dot{P} \{H_1 \hat{B} + H_3\} = [A(Y - \hat{B})]$ ,  $\dot{B} \{H_1 \hat{P} + H_2\} = (\frac{A}{\hat{\pi}} - 1)$ , so

$$\begin{aligned} \widehat{IF}_{22, \tilde{\psi}_k, \tilde{i}_2} &= - \left[ \left( \frac{A}{\hat{\pi}} - 1 \right) \bar{Z}_k^T \right]_{i_1} \left[ \bar{Z}_k A (Y - \hat{B}) \right]_{i_2} \\ \widehat{IF}_{jj, \tilde{\psi}_k, \tilde{i}_j} &= (-1)^{j-1} \left[ \left( \frac{A}{\hat{\pi}} - 1 \right) \bar{Z}_k^T \right]_{i_1} \left[ \prod_{s=3}^j \left\{ A \bar{Z}_k \bar{Z}_k^T - I_{k \times k} \right\}_{i_s} \right] \left[ \bar{Z}_k A (Y - \hat{B}) \right]_{i_2} \end{aligned}$$

Our next theorem, proved in the appendix, derives the estimation bias  $EB_m =$   
 $E[\hat{\psi}_{m,k}] - \tilde{\psi}_k$ .

**Theorem 23** Suppose (21) – (23) and Ai) – Aiv) hold then

$$EB_m = (-1)^{m-1} \left\{ \begin{aligned} & E \left[ Q^2 \left( \frac{B-\hat{B}}{\hat{B}} \right) \bar{Z}_k^T \right] \left\{ E \left[ Q^2 \bar{Z}_k \bar{Z}_k^T \right] - I_{k \times k} \right\}^{m-1} \\ & \times \left\{ E \left[ Q^2 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} E \left[ \bar{Z}_k Q^2 \left( \frac{P-\hat{P}}{\hat{P}} \right) \right] \end{aligned} \right\} \quad (32)$$

$$|EB_m|$$

$$\leq \left\{ \begin{aligned} & \left\| \left\{ \frac{\dot{B}}{\hat{P}} G \right\}^{1/2} \right\|_{\infty} \left\| \left\{ \frac{\dot{P}}{\hat{B}} G \right\}^{1/2} \right\|_{\infty} \|\delta g\|_{\infty}^{m-1} (1 + o_p(1)) \times \\ & \left\{ \int (p(X) - \hat{p}(X))^2 dX \right\}^{1/2} \left\{ \int (b(X) - \hat{b}(X))^2 dX \right\}^{1/2} \end{aligned} \right\} \quad (33)$$

$$= O_P \left( \left( \frac{\log n}{n} \right)^{\frac{(m-1)\beta_g}{d+2\beta_g}} n^{-\left( \frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p} \right)} \right) \quad (34)$$

for  $m \geq 1$ , where  $\delta g = \frac{g(X) - \hat{g}(X)}{\hat{g}(X)}$

**Remark 24** At the cost of a longer proof we could have used Hölder's inequality re-

peatedly to control  $\delta g$  in the  $L_p$  norm  $\|\delta g\|_{m+1}$  with  $p = m + 1$  to show that  $|EB_m| =$

$$O_P \left( \|\delta g\|_{m+1}^{m-1} \left\| b(\cdot) - \hat{b}(\cdot) \right\|_{m+1} \left\| p(\cdot) - \hat{p}(\cdot) \right\|_{m+1} \right). \text{ Thus, } |EB_m| \text{ is } O_P \left( \left\| \theta - \hat{\theta} \right\|^{m+1} \right),$$

consistent with the form of the bias given in our fundamental theorem 2.

**Remark 25** *An alternate derivation of  $\hat{\psi}_{m,k}$ .* The above derivation of  $\hat{\psi}_{m,k}$  re-

quired that one have facility in calculating higher order influence functions  $\mathbb{IF}_{m,\tilde{\psi}_k}$ , as

done in the proof of Theorem 20 in the appendix. However, there exists an alternate

derivation of  $\hat{\psi}_{m,k}$  that does not require one learn how to calculate influence functions.

Specifically, we know from Theorems 2 and 3 that in a (locally) nonparametric model

$\widehat{\mathbb{IF}}_{jj, \tilde{\psi}_k}, j \geq 2$  is the unique  $j^{th}$  order  $U$ -statistic that is degenerate under  $\widehat{\theta}$  and satisfies

$$EB_{j-1} + E \left[ \widehat{\mathbb{IF}}_{jj, \tilde{\psi}_k} | \widehat{\theta} \right] \equiv EB_j = O_p \left( \left\| \widehat{\theta} - \theta \right\|^{j+1} \right) \quad (35)$$

with  $EB_1 = E \left[ \widehat{\psi}_1 | \widehat{\theta} \right] - \tilde{\psi}_k$ . In fact, we first derived  $\widehat{\psi}_{m,k}$  by beginning with  $\widehat{\psi}_1 = \widehat{\psi} + \widehat{\mathbb{IF}}_{1, \tilde{\psi}_k}$ , calculating  $EB_1 = E \left[ \widehat{\psi}_1 | \widehat{\theta} \right] - \tilde{\psi}_k$ , and then, recursively for  $j = 2, \dots$ , finding  $\widehat{\mathbb{IF}}_{jj, \tilde{\psi}_k}$  satisfying the above equation. For explicit details see the Appendix. In fact if one did not even know how to derive  $\widehat{\mathbb{IF}}_{1, \tilde{\psi}_k}$ , one could begin the recursion by obtaining  $\widehat{\mathbb{IF}}_{1, \tilde{\psi}_k}$  as the unique first order  $U$ -statistic with mean zero under  $\widehat{\theta}$  satisfying  $\widehat{\psi} - \tilde{\psi}_k + E \left[ \widehat{\mathbb{IF}}_{1, \tilde{\psi}_k} | \widehat{\theta} \right] = O_p \left( \left\| \widehat{\theta} - \theta \right\|^2 \right)$ .

### 3.2.4 The Variance of $\widehat{\psi}_{m,k} \equiv \widehat{\psi} + \widehat{\mathbb{IF}}_{m, \tilde{\psi}_k}$ using compact wavelets

In this section, we derive the order of the variance of  $\widehat{\psi}_{m,k}$  when the orthonormal system  $\{\varphi_j(X)\}$  used to construct our  $U$ -statistics are a compact wavelet basis. First consider the case where  $X$  is univariate; without loss of generality, assume that  $X \sim \text{Uniform}[0, 1]$ . Because we are primarily interested in convergence rates, the fact that  $X$  may not follow the uniform distribution will not affect the rate results given below, but can influence the size of the constants. We use  $\phi_j(X)$  in place of  $\varphi_j(X)$  to indicate univariate basis functions.

Let  $k^*, k$  be integer powers of two with  $k > k^*$ . Denote by  $\bar{\phi}(X) \equiv \bar{\phi}_1^k(X)$  the  $k$ -dimensional basis vector whose first  $k^*$  components  $\bar{\phi}_1^{k^*}(X)$  are the  $k^*$ -vector of level  $\log_2 k^*$  scaled and translated versions of a compactly supported 'father' wavelet

(Mallat, 1998) and whose last  $k - k^*$  components  $\overline{\phi}_{k^*+1}^k(X)$  are the associated compact mother wavelets between levels  $\log_2 k^*$  and  $\log_2 k$ . In particular, one may use periodic wavelets, folded wavelets or Daubechies' boundary wavelets with enough vanishing moments to obtain the optimal approximation rate of  $O(k^{-2\beta/d})$  for  $\beta = \max(\beta_g, \beta_p, \beta_b)$ . The multiresolution analysis (MRA) property of wavelets allows us to decompose the vector space spanned by the  $\log_2(k)$ -level father wavelets  $\mathcal{V}_{\log_2(k)}$  into the direct sum of the subspace spanned by  $\log_2(k^*)$ -level father wavelets  $\mathcal{V}_{\log_2(k^*)} = \{a^T \overline{\phi}_1^{k^*}(X) : a \in R^{k^*}\}$  and the span of mother wavelets for each level between  $\log_2(k^*)$  and  $\log_2(k) - 1$  which we respectively write as

$$\begin{aligned}\mathcal{W}_{\log_2(k^*)} &= \left\{ a^T \overline{\phi}_{k^*+1}^{2k^*}(X) : a \in R^{k^*} \right\}, \\ \mathcal{W}_{\log_2(k_0)+1} &= \left\{ a^T \overline{\phi}_{2k^*+1}^{4k^*}(X) : a \in R^{2k^*} \right\}, \\ &\vdots \\ \mathcal{W}_{\log_2(k)-1} &= \left\{ a^T \overline{\phi}_{\frac{k}{2}+1}^k(X) : a \in R^{\frac{k}{2}} \right\}.\end{aligned}$$

Then for any integer  $s$  with  $\log_2(k^*) + 1 \leq s$ , we have

$$\mathcal{V}_s = \mathcal{V}_{\log_2(k^*)} \oplus \left( \bigoplus_{v=\log_2(k^*)}^{s-1} \mathcal{W}_v \right)$$

As  $s \rightarrow \infty$ , the resulting basis system is dense in  $L_2(X)$  (Mallat, 1998). Since, in fact,  $X$  is  $d$ -dimensional we require a generalization that allows for multivariate tensor wavelet basis functions. In fact, suppose  $X^T = (X^1, \dots, X^d)$  is now multivariate, and

we again assume  $X \sim \text{Uniform}$  on  $[0, 1]^d$ . Given  $d$  univariate vector spaces

$$\mathcal{V}_{1, \log_2(k)}, \mathcal{V}_{2, \log_2(k)}, \dots, \mathcal{V}_{d, \log_2(k)}$$

respectively spanned by vectors  $\bar{\phi}_1^k(X^1), \bar{\phi}_1^k(X^2), \dots, \bar{\phi}_1^k(X^d)$ , so that for  $1 \leq r \leq d$ ,

$$\mathcal{V}_{r, \log_2(k^*)} \subset \mathcal{V}_{r, \log_2(k^*)+1} \subset \dots \subset \mathcal{V}_{r, \log_2(k)-1} \subset \mathcal{V}_{r, \log_2(k)}$$

and

$$\mathcal{V}_{r, \log_2(k)} = \mathcal{V}_{r, \log_2(k^*)} \oplus \left( \bigoplus_{v=\log_2(k^*)}^{\log_2(k)-1} \mathcal{W}_{r,v} \right)$$

One may define  $d$  dimensional tensor vector spaces

$$\mathcal{Y}_{d, \log_2(k^*)}, \mathcal{Y}_{d, \log_2(k^*)+1}, \dots, \mathcal{Y}_{d, \log_2(k)}$$

such that

$$\mathcal{Y}_{d, \log_2(k^*)} \subset \mathcal{Y}_{d, \log_2(k^*)+1} \subset \dots \subset \mathcal{Y}_{d, \log_2(k)}$$

where for  $s \geq 0$ ,

$$\mathcal{Y}_{d, \log_2(k_0)+s} = \bigotimes_{1 \leq r \leq d} \mathcal{V}_{r, \log_2(k_0)+s}$$

As  $s \rightarrow \infty$ , the resulting tensor basis system is dense in  $L_2(X)$  (Mallat, 1998).

Next, suppose that we have a set of multivariate basis functions

$$\left\{ \bar{\varphi}_1^{k_j}(X), j = 0, 1, \dots, 2m \right\}$$

such that for each  $k_j$ ,  $\bar{\varphi}_1^{k_j}(X)$  spans  $\bigotimes_{1 \leq r \leq d} \mathcal{V}_{r, \log_2(k_{j,r})}$  where  $\prod_{r=1}^d k_{j,r} = k_j$ . Define  $\|\cdot\|_2$

as the  $L_2$  norm with respect to the Lebesgue measure. The following theorem is key

to our derivation of the order of the variance of  $\hat{\psi}_{m,k}$



**Theorem 26** For  $m \geq 0$ ,

$$\begin{aligned}
& \left\| \overline{\varphi}_1^{k_1} (X_{i_1})^T \prod_{j=1}^m \left\{ \overline{\varphi}_1^{k_j} (X_{i_{j+1}}) \overline{\varphi}_1^{k_{j+1}} (X_{i_{j+1}})^T \right\} \overline{\varphi}_1^{k_{m+1}} (X_{i_{m+2}}) \right\|_2^2 \\
&= E \left( \prod_{j=1}^{m+1} K_{(1,k_j)} (X_{i_j}, X_{i_{j+1}}) \right)^2 \\
&\asymp \prod_{j=1}^{m+1} k_j
\end{aligned}$$

The following theorem is an immediate consequence of Theorem (26) obtained by taking  $k_j = k^* = k$  (which implies we use the father wavelets at level  $\log_2(k)$  but no mother wavelets.)

**Theorem 27** For all  $\theta \in \Theta$ ,

$$\begin{aligned}
Var_\theta \left[ \mathbb{IF}_{1, \tilde{\psi}_k} (\theta) \right] &\asymp \frac{1}{n} \\
Var_\theta \left[ \mathbb{IF}_{jj, \tilde{\psi}_k} (\theta) \right] &\asymp \left( \frac{1}{n} \max \left\{ 1, \left( \frac{k}{n} \right)^{j-1} \right\} \right)
\end{aligned}$$

,

$$Var_\theta \left[ \mathbb{IF}_{m, \tilde{\psi}_k} (\theta) \right] \approx Var_{\hat{\theta}} \left[ \widehat{\mathbb{IF}}_{m, \tilde{\psi}_k} | \hat{\theta} \right] \asymp \frac{1}{n} \max \left\{ 1, \left( \frac{k}{n} \right)^{m-1} \right\}$$

We now use Theorem (27) to derive the order of the conditional variance of  $\widehat{\psi}_{m,k}$  given  $\hat{\theta}$ .

**Theorem 28** *If  $\sup_{o \in \mathcal{O}} \left| f(o; \hat{\theta}) - f(o; \theta) \right| \rightarrow 0$  as  $\|\hat{\theta} - \theta\| \rightarrow 0$ , then for a fixed  $m$ ,*

$$\begin{aligned} \text{Var}_{\theta} \left[ \hat{\psi}_{m, \tilde{\psi}_k} | \hat{\theta} \right] &= \text{Var}_{\theta} \left[ \hat{\mathbb{I}\mathbb{F}}_{m, \tilde{\psi}_k} | \hat{\theta} \right] \\ &= \text{Var}_{\hat{\theta}} \left[ \hat{\mathbb{I}\mathbb{F}}_{m, \tilde{\psi}_k} | \hat{\theta} \right] (1 + o_p(1)) \\ &\asymp \left( \frac{1}{n} \max \left\{ 1, \left( \frac{k}{n} \right)^{m-1} \right\} \right) \end{aligned}$$

The proof is in the appendix.

For a given  $m$ , the estimator  $\hat{\psi}_{m, k_{opt}(m)}$  that minimizes the maximum asymptotic MSE over the model  $\mathcal{M}(\Theta)$  defined by  $Ai) - Aiv)$  among the candidates  $\hat{\psi}_{m, k}$  uses the value  $k_{opt}(m) \equiv k_{opt}(m, n)$  of  $k$  that equates the order  $\frac{1}{n} \max \left\{ 1, \left( \frac{k}{n} \right)^{m-1} \right\}$  of  $\text{Var} \left[ \hat{\psi}_{m, \tilde{\psi}_k} | \hat{\theta} \right]$  to the order

$$\begin{aligned} &\max \left[ \{TB_k\}^2, \{EB_m(\theta)\}^2 \right] = \\ &\max \left[ \left( \frac{\log n}{n} \right)^{\frac{2(m-1)\beta_g}{d+2\beta_g}} n^{-\left( \frac{2\beta_b}{d+2\beta_b} + \frac{2\beta_p}{d+2\beta_p} \right)}, \right. \\ &\quad \left. k^{-\frac{2(\beta_b+\beta_p)}{d}} \right] \end{aligned}$$

of the maximal squared bias. The estimator  $\hat{\psi}_{m_{opt}, k_{opt}} \equiv \hat{\psi}_{m_{opt}, k_{opt}(m_{opt})}$  that minimizes the maximum asymptotic MSE over the model  $\mathcal{M}(\Theta)$  among all candidates  $\hat{\psi}_{m, k}$  is the estimator  $\hat{\psi}_{m, k_{opt}(m, n)}$  which minimizes  $\frac{1}{n} \max \left( 1, \left( \frac{k_{opt}(m, n)}{n} \right)^{m-1} \right)$ .

### 3.2.5 Distribution Theory and Confidence Interval Construction

We derive a consistent estimator of the variance and give the asymptotic distribution of  $\hat{\psi}_{m, k}$  for any model and functional satisfying  $Ai) - Aiv)$ . Let  $z_{\alpha}$  be the upper

$\alpha$ -quantile of a standard normal, i.e. a  $N(0, 1)$ , distribution.

**Theorem 29 :**

$$a) \text{ Letting } \widehat{\mathbb{W}}_{1, \tilde{\psi}_k}^2 = n^{-1} \mathbb{V} \left[ \left\{ \widehat{IF}_{1, \tilde{\psi}_k, i_1} \right\}^2 \right],$$

$$\widehat{\mathbb{W}}_{jj, \tilde{\psi}_k}^2 = \binom{n}{j}^{-1} \mathbb{V} \left[ \left( \widehat{IF}_{j, j, \tilde{\psi}_k(\cdot)}^{(s)} \right)^2 \right],$$

for  $j \geq 2$ , and

$$\widehat{\mathbb{W}}_{m, \tilde{\psi}_k}^2 = \widehat{\mathbb{W}}_{1, \tilde{\psi}_k}^2 + \sum_{j=2}^m \widehat{\mathbb{W}}_{jj, \tilde{\psi}_k}^2,$$

where  $\widehat{IF}_{j, j, \tilde{\psi}_k(\cdot)}^{(s)}$  is the symmetric kernel of  $\widehat{\mathbb{IF}}_{jj, \tilde{\psi}_k(\cdot)}$ .

we have,

$$\begin{aligned} \widehat{E} \left[ \widehat{\mathbb{W}}_{1, \tilde{\psi}_k}^2 \right] &= \widehat{Var} \left[ \widehat{\mathbb{IF}}_{1, \tilde{\psi}_k} | \widehat{\theta} \right] \\ \widehat{E} \left[ \widehat{\mathbb{W}}_{jj, \tilde{\psi}_k}^2 \right] &= \widehat{Var} \left[ \widehat{\mathbb{IF}}_{jj, \tilde{\psi}_k} | \widehat{\theta} \right], \\ \widehat{E} \left[ \widehat{\mathbb{W}}_{m, \tilde{\psi}_k}^2 \right] &= \widehat{Var} \left[ \widehat{\mathbb{IF}}_{m, \tilde{\psi}_k} | \widehat{\theta} \right] \end{aligned}$$

where  $\widehat{Var}[\cdot] = Var_{\widehat{\theta}}[\cdot]$ .

b) Conditional on the training sample,

$$\left\{ \frac{1}{n} \max \left\{ 1, \left( \frac{k_{opt}(m, n)}{n} \right)^{m-1} \right\} \right\}^{-1/2} \left\{ \widehat{\psi}_{m, k_{opt}(m, n)} - E \left[ \widehat{\psi}_{m, k_{opt}(m, n)} | \widehat{\theta} \right] \right\}$$

converges uniformly for  $\theta \in \Theta$  to a normal distribution with finite variance as  $n \rightarrow \infty$ .

The asymptotic variance is uniformly consistently estimated by

$$\left\{ \frac{1}{n} \max \left\{ 1, \left( \frac{k}{n} \right)^{m-1} \right\} \right\}^{-1} \widehat{\mathbb{W}}_{m, \tilde{\psi}_{k_{opt}(m, n)}}^2$$

Thus

$$\left\{ \widehat{\psi}_{m,k_{opt}(m,n)} - E \left[ \widehat{\psi}_{m,k_{opt}(m,n)} | \widehat{\theta} \right] \right\} / \widehat{\mathbb{W}}_{m,\widetilde{\psi}_{k_{opt}(m,n)}}$$

is converging in distribution to a standard normal distribution.

c) Define the interval  $C_{m,k} = \widehat{\psi}_{m,k} \pm z_{\alpha} \widehat{\mathbb{W}}_{m,\widetilde{\psi}_k}$ . Suppose  $k_{opt}(m,n) = n^{\rho_{opt}(m,n)}$ .

Then for  $k^* = n^{\rho^*}$ ,  $\rho^* > \rho_{opt}(m,n)$ ,

$$\sup_{\theta \in \Theta} \left[ \frac{E_{\theta} \left[ \widehat{\psi}_{2,k^*} | \widehat{\theta} \right]}{\sqrt{Var_{\theta} \left[ \widehat{\psi}_{2,k^*} | \widehat{\theta} \right]}} \right] = o_p(1)$$

and  $\left\{ \widehat{\psi}_{m,k^*} - \psi(\theta) \right\} / \widehat{\mathbb{W}}_{m,\widetilde{\psi}_{k^*}}$  converges uniformly in  $\theta \in \Theta$  to a  $N(0,1)$ . Moreover,

$C_{m,k^*}$  is a conservative uniform asymptotic  $(1 - \alpha)$  confidence interval for  $\psi(\theta)$ .

d) Suppose we could derive a constant  $C_{bias}$  and a constant  $N^*$  such that

$$\begin{aligned} & \sup_{\theta} \left| E_{\theta} \left[ \left\{ \widehat{\psi}_{m,k_{opt}(m,n)} - \psi(\theta) \right\} \right] \right| \\ &= \sup_{\theta} \left| \{ TB_{k_{opt}(m,n)}(\theta) + EB_m(\theta) \} \right| \\ &\leq C_{bias} \left\{ \frac{1}{n} \max \left\{ 1, \left( \frac{n^{\rho_{opt}(m,n)}}{n} \right)^{m-1} \right\} \right\}^{1/2} \end{aligned}$$

for  $n > N^*$ . Then

$$\begin{aligned} & BC_{m,k_{opt}(m,n)} \\ &= \widehat{\psi}_{m,k_{opt}(m,n)} \pm \left\{ z_{\alpha} \widehat{\mathbb{W}}_{m,\widetilde{\psi}_{k^*}} + C_{bias} \left\{ \frac{1}{n} \max \left\{ 1, \left( \frac{n^{\rho_{opt}(m,n)}}{n} \right)^{m-1} \right\} \right\}^{1/2} \right\} \end{aligned}$$

is a conservative uniform asymptotic  $(1 - \alpha)$  confidence interval for  $\psi(\theta)$ .

Part a) of the theorem is an easy calculation. The asymptotic normality of  $\widehat{\psi}_{m,k_{opt}(m,n)}$  is based on new results on the asymptotic distribution of higher order  $U - statistics$  with kernels depending on  $n$  to be published elsewhere (Robins et al, 2007).

Part c of the theorem implies we obtain a conservative uniform asymptotic  $(1 - \alpha)$  confidence interval for any value of  $\rho^*$  exceeding  $\rho_{opt}(m,n)$ . However, for the actual fixed sample size of our study, say  $n = 5000$ , there is no guarantee the interval of part c based on given difference  $\rho^* - \rho_{opt}(m,n)$ , say .3, will provide conservative finite sample coverage.

Because of this difficulty, a better approach, described in part d, would be to determine a constant  $C_{bias}$  that can be used to bound the maximal bias under the model at a sample sizes exceeding  $N^*$ , with  $N^*$  no greater than the actual fixed sample size  $n$  of the study. Then the interval  $BC_{m,k_{opt}(m,n)}$  will be a honest conservative finite sample  $1 - \alpha$  confidence interval, provided that  $\widehat{\psi}_{m,k_{opt}(m,n)}$  has nearly converged to its normal limit at sample size  $n$ . Unfortunately as yet we do not know how to determine the constants  $C_{bias}$  and  $N^*$  of part d as a function of our model and of our initial estimator  $\widehat{\theta}$ . This is an important open problem.

### 3.2.6 Models of Increasing Dimension and Multi-Robustness

**A Model of Increasing Dimension:** The previous results can also be used for the analysis of models whose dimension increases with sample size. In fact, consider the  $\mathcal{M}(\Theta_{n^\eta})$ ,  $\eta$  known, that differs from model  $\mathcal{M}(\Theta)$  in that, rather than assuming  $b(x)$  and  $p(x)$  live in particular Hölder balls, we instead assume the working models of Eqs. 24 and 25 are precisely true for  $k = n^\eta$ , so  $\psi(\theta) \equiv \tilde{\psi}_{n^\eta}(\theta)$  and the dimensions of  $b(x)$  and  $p(x)$  increase as  $n^\eta$ . Valid point and interval estimation for  $\tilde{\psi}_{n^\eta}(\theta)$  can still be based on the estimators  $\hat{\psi}_{m,k}$  except now (i) there is truncation bias only when  $k < n^\eta$ , (ii) the variance remains of the order of  $\frac{1}{n} \max\left(1, \left(\frac{k}{n}\right)^{m-1}\right)$ , and (iii) the estimation and truncation bias (when it exists) orders will be determined by any additional complexity reducing restrictions placed on the fraction of non-zero components or on the rate of decay of the components of the vectors  $\tilde{\eta}_{n^\eta}(\theta)$  and  $\tilde{\alpha}_{n^\eta}(\theta)$ , and, for estimation bias, by  $\beta_g$  as well. As a consequence,  $m_{opt}$  and  $k_{opt}$  under model  $\mathcal{M}(\Theta_{n^\eta})$  will differ from their values under model  $\mathcal{M}(\Theta)$ . Note we need not take  $k = n^\eta$  as we did in the heuristic discussion following Remark 8. Indeed  $\hat{\psi}_{m_{best}}$  in that discussion corresponds to the estimator in the class  $\hat{\psi}_{m,k=n^\eta}$  with the fastest rate of convergence. In general,  $\hat{\psi}_{m_{best}}$  will have convergence rate slower than  $\hat{\psi}_{m_{opt},k_{opt}}$ . Furthermore, the discussion in Section 4.1.1 implies that, when  $n^\eta \gg n$  and the minimax rate for estimation of  $\psi(\theta)$  is slower than  $n^{-1/2}$ , even  $\hat{\psi}_{m_{opt},k_{opt}}$  will typically fail to converge at the minimax rate when complexity reducing restrictions have been

imposed on  $\widetilde{\eta}_{n^\gamma}(\theta)$  and  $\widetilde{\alpha}_{n^\gamma}(\theta)$ .

**Multi-Robustness and a Practical Data Analysis Strategy:** Conditional on  $\widehat{\theta}$ , for  $m \geq 2$ ,  $EB_m$  is zero and thus estimator  $\widehat{\psi}_{m,k}$  is unbiased for  $\widetilde{\psi}_k$  if  $\widehat{p}(\cdot) = p(\cdot)$ ,  $\widehat{b}(\cdot) = b(\cdot)$ , or  $\widehat{g}(\cdot) = g(\cdot)$ . We refer to  $\widehat{\psi}_{m,k}$  as triply-robust for  $\widetilde{\psi}_k$ , generalizing Robins and Rotnitzky (2001) and van der Laan and Robins (2003) who referred to  $\widehat{\psi}_1$  as doubly-robust because of its being unbiased for  $\widetilde{\psi}_k$  if either  $\widehat{p}(\cdot) = p(\cdot)$  or  $\widehat{b}(\cdot) = b(\cdot)$ . In fact, for  $m \geq 3$ , we can construct a modified estimator  $\widehat{\psi}_{m,k}^{\text{mod}}$  that is  $m+1$ -fold robust as follows. Let  $\widehat{g}_s(\cdot)$ ,  $s = 3, \dots, m$ , denote  $m-2$  additional initial estimators of  $g(\cdot)$  that differ from one another and from  $\widehat{g}(\cdot)$ . Define  $\widehat{\psi}_{m,k}^{\text{mod}} = \widehat{\psi} + \widehat{\mathbb{IF}}_{1,\widetilde{\psi}_k} + \widehat{\mathbb{IF}}_{22,\widetilde{\psi}_k,\bar{i}_j} + \sum_{j=3}^m \widehat{\mathbb{IF}}_{jj,\widetilde{\psi}_k}^{\text{mod}}$ , where

$$\begin{aligned} \widehat{IF}_{jj,\widetilde{\psi}_k,\bar{i}_j}^{\text{mod}} &= (-1)^{j-1} \left[ \left( H_1 \widehat{P} + H_2 \right) \dot{B} \overline{Z}_k^T \right]_{i_1} \left\{ \left( \dot{P} \dot{B} H_1 \overline{Z}_k \overline{Z}_k^T \right)_{i_2} - I_{k \times k} \right\} \\ &\times \left[ \prod_{s=3}^{j-1} \left\{ \widehat{E}_s \left[ \dot{P} \dot{B} H_1 \overline{Z}_k \overline{Z}_k^T \right] \right\}^{-1} \right. \\ &\quad \left. \left\{ \left( \dot{P} \dot{B} H_1 \overline{Z}_k \overline{Z}_k^T \right)_{i_s} - \widehat{E}_s \left[ \dot{P} \dot{B} H_1 \overline{Z}_k \overline{Z}_k^T \right] \right\} \right] \\ &\times \left\{ \widehat{E}_j \left[ \dot{P} \dot{B} H_1 \overline{Z}_k \overline{Z}_k^T \right] \right\}^{-1} \times \left[ \overline{Z}_k \left( H_1 \widehat{B} + H_3 \right) \dot{P} \right]_{i_j} \end{aligned}$$

with  $\widehat{E}_s$  defined like  $\widehat{E}$ , except with  $\widehat{g}_s(\cdot)$  replacing  $\widehat{g}(\cdot)$ . In the appendix, we prove

that  $EB_m^{\text{mod}} = E \left[ \widehat{\psi}_{m,k}^{\text{mod}} \right] - \widetilde{\psi}_k$  is

$$(-1)^{m-1} \left\{ \begin{aligned} & E \left[ \dot{B} \dot{P} H_1 \left( \frac{P - \widehat{P}}{\widehat{P}} \right) \overline{Z}_k^T \right] \left\{ E \left[ \dot{B} \dot{P} H_1 \overline{Z}_k \overline{Z}_k^T \right] - I_{k \times k} \right\} \\ & \quad \times \prod_{s=3}^m \left\{ \widehat{E}_s \left[ \dot{B} \dot{P} H_1 \overline{Z}_k \overline{Z}_k^T \right] \right\}^{-1} \\ & \quad \times \left\{ E \left[ \dot{B} \dot{P} H_1 \overline{Z}_k \overline{Z}_k^T \right] - \widehat{E}_s \left[ \dot{B} \dot{P} H_1 \overline{Z}_k \overline{Z}_k^T \right] \right\} \\ & \quad \times \left\{ E \left[ \dot{B} \dot{P} H_1 \overline{Z}_k \overline{Z}_k^T \right] \right\}^{-1} E \left[ \dot{B} \dot{P} H_1 \left( \frac{B - \widehat{B}}{\widehat{B}} \right) \right] \end{aligned} \right\} \quad (36)$$

which is zero if  $\widehat{p}(\cdot) = p(\cdot)$ ,  $\widehat{b}(\cdot) = b(\cdot)$ ,  $\widehat{g}(\cdot) = g(\cdot)$ , or if any of the  $m - 2$   $\widehat{g}_s(\cdot)$  equals  $g(\cdot)$ . (We note that if  $\widehat{p}(\cdot) = p(\cdot)$  or  $\widehat{b}(\cdot) = b(\cdot)$ ,  $\psi = \widetilde{\psi}_k$  and thus  $\widehat{\psi}_{m,k}^{\text{mod}}$  and  $\widehat{\psi}_{m,k}$  are unbiased for  $\psi$ .)

In settings where the dimension  $d$  of  $X$  is so large (say  $30 - 50$ ) that the above asymptotic results fail as a guide to the finite sample performance of our procedures at the moderate sample sizes, say  $n = 500 - 5000$ , commonly found in practice, one might consider, as a practical data analysis strategy, using the  $m + 1 - fold$  robust estimator  $\widehat{\psi}_{m,k}^{\text{mod}}$  with  $\widehat{p}(\cdot)$ ,  $\widehat{b}(\cdot)$ ,  $\widehat{g}(\cdot)$ , and the  $\widehat{g}_s(\cdot)$  selected by cross-validation as in van der Laan and Dudoit (2003). Specifically, the training sample is split into two random subsamples - a candidate estimator subsample of size  $n_c$  and a validation subsample of size  $n_v$ , where both  $n_c/n$  and  $n_v/n$  are bounded away from 0 as  $n \rightarrow \infty$ . A large number (e.g.,  $n^3$ ) candidate parametric models of various dimensions and functional forms for  $p$ ,  $b$ , and  $g$  are fit to the candidate estimator subsample and the validation sample is used to find the candidate estimators  $\widehat{p}(\cdot)$  and  $\widehat{b}(\cdot)$  for  $p$  and  $b$  and the  $m - 1$  candidate estimators  $\widehat{g}(\cdot)$  and  $\widehat{g}_s(\cdot)$ ,  $s = 3, \dots, m$ , for  $g$  with the



smallest estimated risks (with respect to an appropriate risk function such as squared error or Kullback-Leibler.) An alternative approach would be to use the triply robust estimator  $\widehat{\psi}_{m,k}$  with  $\widehat{g}(\cdot)$  the candidate for  $g$  with minimum estimated risk. We plan to explore through simulation whether  $\widehat{\psi}_{m,k}^{\text{mod}}$  outperforms  $\widehat{\psi}_{m,k}$  in the setting of very high dimensional  $X$ .

#### 4 Rates of Convergence and Minimavity

We consider a generic version in which we only assume a model and functional satisfying  $Ai) - Aiv)$ . To examine efficiency issues, we first consider the estimator  $\widehat{\psi}_1$  based on the first order influence function and sample splitting. Without loss of generality we assume  $\beta_p \geq \beta_b$ . (Otherwise simply interchange  $\beta_p$  and  $\beta_b$  in what follows.) It will be useful to consider the alternative parametrization

$$\beta = \frac{\beta_p + \beta_b}{2},$$

$$\Delta = \left( \frac{\beta_p}{\beta_b} - 1 \right) \geq 0$$

The (conditional) variance of  $\widehat{\psi}_1$  is of the order of  $1/n$  and the (conditional) bias of  $\widehat{\psi}_1$  in estimating  $\psi$  is  $O_p \left( n^{-\left( \frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p} \right)} \right)$ . If  $\Delta = 0$  and thus  $\beta_p = \beta_b$ , the bias of  $\widehat{\psi}_1$  is  $n^{-\frac{2\beta}{d+2\beta}}$  and  $\widehat{\psi}_1$  is not  $n^{1/2}$ -consistent for  $\psi$  when  $\beta/d < 1/2$ . At the other extreme, as  $\Delta \rightarrow \infty$ , i.e.  $\beta_b \rightarrow 0$ , the bias of  $\widehat{\psi}_1$  is  $n^{-\frac{2\beta}{d+4\beta}}$  which fails to be  $n^{1/2}$ -consistent for any finite  $\beta$ .

**Minimaxity with  $g$  known:** To further examine efficiency issues, it is instructive to first consider the estimation of  $\psi$  with  $g(\cdot)$  known. If  $g(\cdot)$  were known, we could set  $\widehat{g}(X) = g(X)$  when calculating  $\widehat{\psi}_{m,k}$ . Then  $EB_2 = 0$  and  $\widehat{\psi}_{2,k}$  would therefore be an unbiased estimator of  $\widetilde{\psi}_k$ . Letting a superscript  $g$  denote the model with  $g$  known, it is easy to see that  $\widehat{\psi}_{m_{opt}^g, k_{opt}^g(m_{opt}^g)}$  would be  $\widehat{\psi}_{2, k_{opt}^g(2)}$  where  $k_{opt}^g(2)$  satisfies  $\max(1/n, k/n^2) \asymp \text{Var}(\widehat{\psi}_{2,k}) = TB_k^2 = k^{-2(\beta_b + \beta_p)/d} = k^{-4\beta/d}$ . Solving this, we find that when  $\beta/d$  is greater than or equal to  $1/4$ , we can take  $k_{opt}^g = n^{\frac{1}{4\beta/d}} \leq n$  and  $|\widehat{\psi}_{2, k_{opt}^g(2)} - \psi| = O_p(n^{-\frac{1}{2}})$  regardless of  $\Delta$ , which is, of course, the minimax rate.

In contrast if  $\beta/d < 1/4$ ,  $k_{opt}^g(2) = n^{\frac{2}{1+4\beta/d}}$  and  $|\widehat{\psi}_{2, k_{opt}^g(2)} - \psi| = n^{-\frac{4\beta}{4\beta+d}}$ . In an unpublished paper, we have proved that this is the minimax rate when  $g(\cdot)$  is known.

This raises the question of whether the lower bounds of rate  $n^{-\frac{1}{2}}$  for  $\beta/d \geq 1/4$  and/or rate  $n^{-\frac{4\beta}{4\beta+d}}$  for  $\beta/d < 1/4$  are still achievable when  $g$  is unknown, without restrictions on the smoothness of  $g$ .

Before addressing this question, we take the opportunity to compare the relative efficiencies of competing rate-optimal unbiased estimators in the case of  $g$  known. This discussion will provide further insight into the results given in Remark 6 for models which are not locally nonparametric.

### **Relative Efficiency of Various Unbiased Estimator with $g$ known:**

For simplicity, we restrict the following discussion to the truncated version of the parameter  $\psi = E[\{b(X)\}^2]$ , with  $b(X) = E[Y|X]$ ,  $g(\cdot)$  known, and  $Y$  Bernoulli.

For this choice of  $\psi$ ,  $g(\cdot)$  is the marginal density of  $X$ . In this subsection, we assume  $\hat{g}(X)$  is chosen equal to the known  $g(X)$  so  $E[\bar{Z}_k \bar{Z}_k^T] = I_{k \times k}$ . Also we choose  $\dot{B} = \dot{P} = 1$  and take  $\hat{B} = \hat{b}(X) \in \text{lin}\{\bar{Z}_k\}$ , so  $\tilde{B} = \Pi[B|\bar{Z}_k] = E[B\bar{Z}_k^T]\bar{Z}_k$  and  $\tilde{\psi}_k = E[\{\Pi[B|\bar{Z}_k]\}^2]$  do not depend on  $\hat{B}$ . Further we only concern ourselves with efficiency relative to the  $n$  observations in the estimation sample. We thus ignore any efficiency loss from using  $N - n$  observations to construct  $\hat{b}$ .

Let  $\Theta^g = \{b : x \mapsto b(x) \in [0, 1]\} \subset \Theta$  denote the subset of  $\Theta$  corresponding to the known  $g$ , which consists of all functions from the unit cube in  $R^d$  to the unit interval. The model  $\mathcal{M}(\Theta^g)$  is not locally nonparametric. For example, the 1st order tangent space  $\Gamma_1(\theta)$  does not include first order scores for  $g$ . Its 2nd order tangent space  $\Gamma_2(b)$  does not contain second order scores for  $g$  or mixed scores for  $g$  and  $b$ . Rather,  $\Gamma_2(b)$  is the closed linear span of the first and second order scores for  $b$ . Thus

$$\Gamma_2(b) = \{\mathbb{S}(a, c); \text{var}_b[\mathbb{S}(a, c)] < \infty; a \in \mathcal{A}, c \in \mathcal{C}\}$$

where

$$S_{ij}(a, c) = \{(Y - B)a(X)\}_i + \left[(Y - B)_i c(X_i, X_j)(Y - B)_j\right],$$

and  $\mathcal{A}$  and  $\mathcal{C}$  are the set of one and two dimensional functions of  $x$ . Since, for  $\hat{b} \in$

$\text{lin} \{ \bar{z}_k(x) \}$

$$\begin{aligned} \hat{\psi}_{2,k}(\hat{b}) &\equiv \hat{\psi}_{2,k} \\ &= \mathbb{V} \left\{ \begin{aligned} &\left[ \hat{B}^2 + 2\hat{B}(Y - \hat{B}) \right]_i \\ &+ \left[ (Y - \hat{B}) \bar{Z}_k^T \right]_i \left[ \bar{Z}_k (Y - \hat{B}) \right]_j \end{aligned} \right\} \end{aligned}$$

is unbiased for  $\tilde{\psi}_k(b) = E \left[ \left\{ \Pi [B|\bar{Z}_k] \right\}^2 \right]$  in model  $\mathcal{M}(\Theta^g)$ , we know, by Remark 6, that  $\mathbb{IF}_{2,\tilde{\psi}_k}^{eff}(b)$  for  $\tilde{\psi}_k(b)$  is the projection  $\Pi_b \left[ \hat{\psi}_{2,k} - \tilde{\psi}_k(b) | \Gamma_2(\theta) \right]$  of the 2nd order influence function  $\hat{\psi}_{2,k} - \tilde{\psi}_k(b)$  onto  $\Gamma_2(b)$ . Now if  $\hat{\psi}_{2,k} - \tilde{\psi}_k(b)$  was an element of  $\Gamma_2(b)$ ,  $\hat{\psi}_{2,k} - \tilde{\psi}_k(b)$  would equal  $\mathbb{IF}_{2,\tilde{\psi}_k}^{eff}(b)$  and thus be 2nd order ‘unbiased locally efficient’, at  $b \in \Theta^g$ , as defined earlier in Remark 6. However we show below that, when  $\hat{b}(X) = c$  for some  $c$  w.p.1 does not hold,  $\hat{\psi}_{2,k} - \tilde{\psi}_k(b)$  is not an element of  $\Gamma_2(b)$  for any  $b$ . Rather, a straightforward calculation gives

$$\mathbb{IF}_{2,\tilde{\psi}_k}^{eff}(b) = \mathbb{V} \left\{ \begin{aligned} &\left[ 2E \left[ B \bar{Z}_k^T \right] \bar{Z}_k (Y - B) \right]_i \\ &+ \left[ (Y - B) \bar{Z}_k^T \right]_i \left[ \bar{Z}_k (Y - B) \right]_j \end{aligned} \right\}.$$

Now one can check that  $\tilde{\psi}_k(\hat{b}) + \mathbb{IF}_{2,\tilde{\psi}_k}^{eff}(\hat{b})$  is a function of  $\hat{b}$ , so by Theorem 7 of Remark 6, we conclude no unbiased globally efficient estimator exists. However, we prove below that  $\tilde{\psi}_k(\hat{b}) + \mathbb{IF}_{2,\tilde{\psi}_k}^{eff}(\hat{b})$  and  $\hat{\psi}_{2,k}$  have identical means. It follows that  $\tilde{\psi}_k(\hat{b}) + \mathbb{IF}_{2,\tilde{\psi}_k}^{eff}(\hat{b})$  is an unbiased estimator of  $\tilde{\psi}_k(b) = E \left[ \left( \Pi [B|\bar{Z}_k] \right)^2 \right]$  for any  $\hat{b} \in \text{lin} \{ \bar{z}_k(x) \}$ . Thus, for a given choice of  $\hat{b} \in \text{lin} \{ \bar{z}_k(x) \}$ ,  $\tilde{\psi}_k(\hat{b}) + \mathbb{IF}_{2,\tilde{\psi}_k}^{eff}(\hat{b})$  is 2nd order unbiased locally efficient at  $b = \hat{b}$ . However, one can show using a proof

analogous to that in theorem 28 that for  $k \ll n^2$

$$\begin{aligned} & var_b \left[ \tilde{\psi}_k(\hat{b}) + \mathbb{IF}_{2, \tilde{\psi}_k}^{eff}(\hat{b}) \right] / var_b \left[ \mathbb{IF}_{2, \tilde{\psi}_k}^{eff}(b) \right] \\ &= 1 + o_P \left( \left\| \hat{b} - b \right\|_\infty \right). \end{aligned}$$

Henceforth we assume that  $b$  lies in a Holder ball  $H(\beta_b, C_b)$ . That is we consider the submodel  $b \in \Theta^g \cap H(\beta_b, C_b)$  and assume  $\hat{b}(x) \in \text{lin} \{ \bar{z}_k(x) \}$  converges to  $b$  in sup norm at the optimal rate of  $\left( \frac{n}{\log n} \right)^{-\beta_b/(2\beta_b+d)}$  uniformly over  $\Theta^g \cap H(\beta_b, C_b)$ . The submodel and the model  $\Theta^g$  have identical tangent spaces. For all  $b \in \Theta^g \cap H(\beta_b, C_b)$ ,  $(\max(n^{-1}, k/n^2))^{-1/2} \left\{ \tilde{\psi}_k(\hat{b}) + \mathbb{IF}_{2, \tilde{\psi}_k}^{eff}(\hat{b}) - \tilde{\psi}_k(b) \right\}$  has an asymptotic distribution with mean zero and variance equal to  $\lim_{n \rightarrow \infty} (\max(n^{-1}, k/n^2))^{-1} var_b \left[ \mathbb{IF}_{2, \tilde{\psi}_k}^{eff}(b) \right]$  for all  $b \in \Theta^g \cap H(\beta_b, C_b)$ . In a slight abuse of language, we shall refer to  $var_b \left[ \mathbb{IF}_{2, \tilde{\psi}_k}^{eff}(b) \right]$  as the asymptotic variance of  $\left\{ \tilde{\psi}_k(\hat{b}) + \mathbb{IF}_{2, \tilde{\psi}_k}^{eff}(\hat{b}) - \tilde{\psi}_k(b) \right\}$ . Thus, as with standard first order theory, even when no unbiased estimator has finite sample variance that attains the Bhattacharyya bound for all  $b \in \Theta^g \cap H(\beta_b, C_b)$ , there can exist an unbiased estimator sequence whose asymptotic variance does attain the bound globally.

We next compare the means and variances of  $\tilde{\psi}_k(\hat{b}) + \mathbb{IF}_{2, \tilde{\psi}_k}^{eff}(\hat{b})$  and  $\hat{\psi}_{2,k}$ . Now the two estimators are algebraically related by

$$\hat{\psi}_{2,k} = \left\{ \tilde{\psi}_k(\hat{b}) + \mathbb{IF}_{2, \tilde{\psi}_k}^{eff}(\hat{b}) \right\} + \left\{ \mathbb{V}[\hat{B}^2] - E[\hat{B}^2] \right\}.$$

Since  $\mathbb{V}[\hat{B}^2] - E[\hat{B}^2]$  is unbiased for zero, we conclude that  $\hat{\psi}_{2,k}$  and  $\tilde{\psi}_k(\hat{b}) + \mathbb{IF}_{2, \tilde{\psi}_k}^{eff}(\hat{b})$  have the same mean but  $var_b \left[ \hat{\psi}_{2,k} \right] / var_b \left[ \mathbb{IF}_{2, \tilde{\psi}_k}^{eff}(b) \right] > 1$  except when

$\widehat{b}(X) = b(X) = c$  wp 1 for some  $c$ . Thus, since  $\widetilde{\psi}_k(\widehat{b}) + \mathbb{IF}_{2, \widetilde{\psi}_k}^{eff}(\widehat{b})$  has asymptotic variance  $var_b \left[ \mathbb{IF}_{2, \widetilde{\psi}_k}^{eff}(b) \right]$  and, except when  $\widehat{b}(X) = c + o_p(1)$ ,  $var \left( \mathbb{V} \left[ \widehat{B}^2 \right] \right) \asymp n^{-1}$ , we conclude the asymptotic variance of  $\widehat{\psi}_{2,k}$  attains the bound  $var_b \left[ \mathbb{IF}_{2, \widetilde{\psi}_k}^{eff}(b) \right]$  when  $k \gg n$ , but exceeds the bound when  $k \leq n$ , except when  $\widehat{b}(X) = c + o_p(1)$ .

Finally, for completeness, Robins and van der Vaart (2006) considered an alternative particularly simple rate-optimal unbiased estimator of  $\widetilde{\psi}_k(b) = E \left[ \left\{ \Pi \left[ B | \overline{Z}_k \right] \right\}^2 \right]$  given by  $\widehat{\psi}_{RV} = \mathbb{V} \left\{ \left[ Y \overline{Z}_k^T \right]_i \left[ \overline{Z}_k Y \right]_j \right\}$ . The Hoeffding decomposition of  $\widehat{\psi}_{RV} - \widetilde{\psi}_k(b)$  is

$$\begin{aligned} & \mathbb{V} \left[ E \left[ B \overline{Z}_k^T \right] \overline{Z}_k Y - \widetilde{\psi}_k(b) \right] + \mathbb{V} \left\{ \left[ Y \overline{Z}_k^T - E \left[ B \overline{Z}_k^T \right] \right]_i \left[ \overline{Z}_k Y - E \left[ B \overline{Z}_k \right] \right]_j \right\} \\ &= \mathbb{IF}_{2, \widetilde{\psi}_k}^{eff}(b) + Q + T \end{aligned}$$

with

$$\begin{aligned} Q &= \mathbb{V} \left[ \left\{ \Pi \left[ B | \overline{Z}_k \right] B - \psi \right\} \right] \\ T &= \mathbb{V} \left\{ \begin{aligned} & 2 \left( B_i \overline{Z}_{k,i}^T \overline{Z}_{k,j} - \Pi \left[ B | \overline{Z}_k \right]_j \right) (Y - B)_j \\ & + B_i \overline{Z}_{k,i}^T \overline{Z}_{k,j} B_j - \Pi \left[ B | \overline{Z}_k \right]_i B_i - \Pi \left[ B | \overline{Z}_k \right]_j B_j + \psi \end{aligned} \right\} \end{aligned}$$

Since, except when  $B = c$  wp1,  $var_b(Q) \asymp n^{-1}$  and  $var_b(T) \asymp k/n^2$ , we conclude that the asymptotic variance of  $\widehat{\psi}_{RV}$  exceeds the bound  $var_b \left[ \mathbb{IF}_{2, \widetilde{\psi}_k}^{eff}(b) \right]$  regardless of whether  $k \gg n$  does or does not hold except when  $b(X) = c$  wp1.

**Minimaxity with Unknown  $g$  and  $\beta/d \geq 1/4$ :** We now show that the bound  $n^{-\frac{1}{2}}$  for  $\beta/d \geq 1/4$  is achievable for each  $\beta_g > 0$ . Consider the estimator  $\widehat{\psi}_{m,k}$  with  $n^{\frac{2}{1+4\beta/d}} \leq k \leq n$  and

$$m \geq 1 + \left\{ \frac{1}{2} - \frac{\beta_b}{d + 2\beta_b} - \frac{\beta_p}{d + 2\beta_p} \right\} \frac{2\beta_g + d}{\beta_g}$$

so that  $EB_m = O_p \left( n^{-\left( \frac{(m-1)\beta_g}{2\beta_g+d} + \frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p} \right)} \right)$  is  $O_p(n^{-1/2})$ . Then  $Var(\widehat{\psi}_{m,k}) \asymp 1/n$ ,  $TB_k^2 = O_p(1/n)$  and  $EB_m^2 = O_p(1/n)$  so  $\widehat{\psi}_{m,k}$  will be  $n^{\frac{1}{2}}$ -consistent for  $\psi$ .

If  $\Delta = 0$  and  $\beta < 1/2$ , the above expression implies that  $m \geq \frac{d-2\beta}{2(2\beta+d)} / \frac{\beta_g}{(2\beta_g+d)} + 1$  for  $n^{\frac{1}{2}}$ -consistency. Similarly, if  $\Delta \rightarrow \infty$ , i.e.  $\beta_b \rightarrow 0$ , it is necessary that  $m \geq \frac{d}{2(4\beta+d)} / \frac{\beta_g}{(2\beta_g+d)} + 1$  for  $n^{\frac{1}{2}}$ -consistency. These results imply that estimators  $\widehat{\psi}_{m,k}$  in our class can always achieve  $n^{\frac{1}{2}}$ -consistency whenever  $\beta_g > 0$ , but for fixed  $\beta < d/2$ , the order  $m$  of the required  $U$ -statistic increases without bound as the smoothness  $\beta_g$  of  $g$  approaches zero.

**Efficiency:** We now show that when  $\beta/d$  is strictly greater than  $1/4$ , we can construct an unconditional asymptotically linear estimator based on all  $N$  subjects with influence function  $N^{-1} \sum_{i=1}^N IF_{1,\psi,i}(\theta)$  by having the number of the  $N$  subjects allotted to the validation sample and analysis sample be  $N^{1-\epsilon}$  and  $n = n(\epsilon) = N - N^{1-\epsilon}$ , respectively, for  $1 > \epsilon > 0$ . It then follows from van der Vaart (1998) that the estimator is regular and semiparametric efficient. Specifically, suppose  $\beta/d = 1/4 + \delta$ ,  $\delta > 0$ . Consider the estimator  $\widehat{\psi}_{m^*,k}$  with  $m^* > 1 + \left\{ \frac{1}{2(1-\epsilon)} - \frac{\beta_b}{d+2\beta_b} - \frac{\beta_p}{d+2\beta_p} \right\} \frac{2\beta_g+d}{\beta_g}$  so

that  $EB_{m^*} = O_p \left( N^{-(1-\epsilon) \left( \frac{(m^*-1)\beta_g}{2\beta_g+d} + \frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p} \right)} \right)$  is  $o_p(N^{-1/2})$  and  $k = n(\epsilon)^{\frac{1}{1+2\delta}} < n(\epsilon)$  so that  $TB_k^2 = o_p(1/N)$  and  $\text{var} \left[ \widehat{IF}_{jj, \tilde{\psi}_k} \right] = o_p(1/N)$  for  $j \geq 2$ . Then, by our previous results,

$$\widehat{\psi}_{m^*,k} - \psi(\theta) = n(\epsilon)^{-1} \sum_{i=1}^{n(\epsilon)} IF_{1,\psi,i}(\theta) + o_p(N^{-1/2}).$$

It remains to show that  $N^{-1} \sum_{i=1}^N IF_{1,\psi,i}(\theta) - n(\epsilon)^{-1} \sum_{i=1}^{n(\epsilon)} IF_{1,\psi,i}(\theta) = o_p(N^{-1/2})$ .

But the LHS is

$$\begin{aligned} & n(\epsilon)^{-1} \sum_{i=1}^{n(\epsilon)} IF_{1,\psi,i}(\theta) \left\{ \frac{n(\epsilon)}{N} - 1 \right\} + N^{-1} \sum_{i=n(\epsilon)+1}^N IF_{1,\psi,i}(\theta) \\ &= O_p \left( n(\epsilon)^{-1/2} N^{-\epsilon} \right) + O_p \left( N^{-(1-\epsilon)/2} N^{-1} \right) = O_p \left( N^{-1/2} N^{-\epsilon} \right) + O_p \left( N^{-1/2} N^{-\epsilon/2} \right) \\ &= o_p(N^{-1/2}). \end{aligned}$$

**Adaptivity when  $\beta/d > 1/4$  :** We next prove that if we let  $n \equiv n(\epsilon) = N - N^{1-\epsilon}$ ,  $m \equiv m(N) = o(N)$  with  $\ln(N) = O(m(N))$  and  $k = n(\epsilon)/\ln(n)$ ,  $\widehat{\psi}_{m,k}$  will be semiparametric efficient for each  $\beta > 1/4$ , provided  $\{\widehat{g}(X) - g(X)\} = o_p \left( m(N^{1-\epsilon})^{-2} \right)$ . Clearly, the truncation bias is  $o(N^{-1/2})$ . The estimation bias  $EB_{m(N)}$  is  $O_p \left( m(N^{1-\epsilon})^{-2[m(N)-1]} N^{-(1-\epsilon) \left\{ \frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p} \right\}} \right)$ . Thus  $EB_{m(N)} = o_p(N^{-1/2})$  if  $m(N^{1-\epsilon})^{-2[m(N)-1]} = o \left( N^{-\frac{1}{2} + (1-\epsilon) \left\{ \frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p} \right\}} \right)$ . So we require

$$2[m(N) - 1] \ln \{m(N^{1-\epsilon})\} / \left[ \frac{1}{2} - (1-\epsilon) \left\{ \frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p} \right\} \right] \ln(N) \rightarrow \infty,$$

which is satisfied if  $\ln(N) = O(m(N))$ . In the appendix we prove that  $\text{var}_\theta \left[ \widehat{\psi}_{m,k} \right] =$

$\text{var}_{\widehat{\theta}} \left[ \widehat{\psi}_{m,k} \right] \{1 + o_p(1)\}$  provided  $\{\widehat{g}(X) - g(X)\} = o_p \left( m(N^{1-\epsilon})^{-2} \right)$ . Now  $\text{var}_{\widehat{\theta}} \left[ \widehat{\psi}_{m,k} \right] =$



$\frac{1}{n} \text{var}_{\hat{\theta}} \left\{ IF_{1,\psi,i}(\hat{\theta}) \right\} \left[ O \left( \sum_{l=0}^{m(N)} \{\ln n\}^{-l} \right) \right]$ . But  $\sum_{l=0}^{m(N)} \{\ln n\}^{-l} = O \left( \frac{1 - (\ln n)^{-[m(N)+1]}}{1 - \{\ln n\}^{-1}} \right) = O \left( 1 + \{\ln n\}^{-1} \right)$ , so  $\text{var}_{\theta} \left[ \hat{\psi}_{m,k} \right]$  is  $n^{-1} \text{var}_{\hat{\theta}} \left\{ IF_{1,\psi,i}(\hat{\theta}) \right\} \{1 + o_p(1)\} = n^{-1} \text{var} \left\{ \mathbb{IF}_{1,\psi} \right\} \{1 + o_p(1)\}$ .

The proof of efficiency now proceeds as above.

**Alternative Estimators when  $\beta/d > 1/4$  :** When  $\beta/d > 1/4$ , there actually exist, at least for certain functionals in our class,  $n^{\frac{1}{2}}$ -consistent estimators of  $\psi$  that are much simpler than our very high order U-statistic estimators. For example consider the expected conditional covariance  $\psi = E[Cov\{A, Y|X\}]$  of Example 1b with  $d = 1$ .

**Example 1b (cont):** Number the study subjects  $i = 0, \dots, N-1$  ordered by their realized values  $X_i$ , where we have not split the sample. Following Wang et al. (2006), consider the difference-based estimator

$$\hat{\psi}_d = N^{-1} \sum_{i=0}^{N/2-1} \{Y_{2i}A_{2i} + Y_{2i+1}A_{2i+1} - Y_{2i+1}A_{2i} - Y_{2i}A_{2i+1}\}$$

which has conditional mean given  $\{X_1, \dots, X_N\}$  of

$$\begin{aligned} & N^{-1} \sum_{i=0}^{N/2-1} Cov\{A, Y|X_{2i}\} + Cov\{A, Y|X_{2i+1}\} \\ & + N^{-1} \sum_{i=0}^{N/2-1} (\{b(X_{i+1}) - b(X_i)\} \{p(X_{i+1}) - p(X_i)\}) \end{aligned}$$

Hence

$$\begin{aligned}
& E \left[ \widehat{\psi}_d - \psi \right] \\
&= N^{-1} E \left[ \sum_{i=0}^{N/2-1} \{b(X_{i+1}) - b(X_i)\} \{p(X_{i+1}) - p(X_i)\} \right] \\
&= O_p \left( N^{-1} \sum_{i=0}^{\frac{N}{2}-1} E \{X_{i+1} - X_i\}^{2\beta} \right) = O(N^{-2\beta})
\end{aligned}$$

by the theory of spacings (Pyke, 1965). But  $O(N^{-2\beta})$  is  $o_p(N^{-1/2})$  when  $\beta > 1/4$ . The variance of  $\widehat{\psi}_d$  is  $O(N^{-1})$  so  $\widehat{\psi}_d$  is  $N^{1/2}$ -consistent. However,  $\frac{\text{var}_\theta(\widehat{\psi}_d)}{\text{var}_\theta(\mathbb{IF}_{1,\psi}(\theta))} \neq 1 + o_p(1)$  so  $\widehat{\psi}_d$  is not (semiparametric) efficient. As discussed by Arellano (2003), by using a  $m$ th order rather than a second order difference operator and letting  $m \rightarrow \infty$  at an appropriate rate as  $N \rightarrow \infty$ , the  $m$ th order estimator  $\widehat{\psi}_d$  can be made efficient.

**Minimaxity with Unknown  $g$  and  $\beta/d < 1/4$  :** Consider next whether the lower bound of  $n^{-\frac{4\beta}{4\beta+d}}$  for  $\beta/d < 1/4$  is achievable when  $g$  is unknown but  $\beta_g > 0$ . We will show the next section that the bound  $n^{-\frac{4\beta}{4\beta+d}}$  is achievable provided

$$\frac{2\beta_g/d}{2\beta_g/d + 1} > \frac{4\beta/d \frac{1-4\beta/d}{1+4\beta/d} (\Delta + 1)}{(\Delta + 2)}, \quad (37)$$

i.e.,  $\beta_g > \frac{2\beta(\Delta+1)(1-4\beta/d)}{(\Delta+2)(1+4\beta/d)-4(\beta/d)(1-4\beta/d)(\Delta+1)}$ . To attain the bound  $n^{-\frac{4\beta}{4\beta+d}}$  whenever eq.(37) holds, we introduce new more efficient estimators, owing to the fact that an estimator  $\widehat{\psi}_{m,k}$  in our class can attain the bound  $n^{-\frac{4\beta}{4\beta+d}}$  only in the special case where the second order estimation bias  $EB_2 = O_p \left( n^{-\left(\frac{\beta_g}{2\beta_g+d} + \frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p}\right)} \right)$  is less than  $n^{-\frac{4\beta}{4\beta+d}}$ .

For a fixed  $\beta = (\beta_p + \beta_b)/2$ , the right hand side of eq.(37) is minimized over  $\Delta \geq 0$  at  $\Delta = 0$ . At  $\Delta = 0$ , eq.(37) reduces to

$$\frac{\beta_g/d}{2\beta_g/d + 1} > \frac{1 - 4\beta/d}{1 + 4\beta/d} \beta/d \Rightarrow \quad (38)$$

$$\beta_g > \frac{\beta(1 - 4\beta/d)}{1 + 2\beta/d + 8(\beta/d)^2} \quad (39)$$

The right hand side of eq.(37) increases with  $\Delta$  with asymptote equal to twice the RHS of eq.(38) as  $\Delta \rightarrow \infty$ . Hence, in order to attain the optimal rate  $n^{-\frac{4\beta}{4\beta+d}}$  when  $\beta_p = 2\beta$  and  $\beta_b = 0$ , the quantity  $\frac{\beta_g}{2\beta_g+d}$  must be twice as large as when  $\beta_p = \beta_b = \beta$ .

In the next section, we construct an estimator with a convergence rate of  $\log(n) n^{-\frac{4\beta}{4\beta+d}}$  at the cut-point  $\frac{\beta_g}{1+2\beta_g} = \frac{(1-4\beta)\beta}{1+4\beta}$ . In this paper we do not consider the construction of estimators that are rate optimal below this cutpoint.

However, for the special case  $\Delta = 0$ , in an unpublished paper Li et. al. (2007) have constructed estimators which converge at a rate given in Eq.(3), whenever inequality (37) fails to hold. We conjecture that this rate is minimax, possibly only up to log factors, when inequality (37) fails to hold and  $\Delta = 0$ . At the cut-point  $\frac{\beta_g}{1+2\beta_g} = \frac{(1-4\beta)\beta}{1+4\beta}$ , we obtain  $m^* = 0$  and thus Eq.(3) becomes  $\log(n) n^{-\frac{4\beta}{4\beta+d}}$ , in agreement with the rate of the estimator of Section 4.1.2 below. In the extreme case in which  $\beta_g \rightarrow 0$  with  $\beta$  remaining fixed,  $\log(n) n^{-\frac{1}{2} + \frac{\beta_g/d}{1+2\beta_g/d} \frac{(m^*+1)^2}{2\beta/d}} \rightarrow \log(n) n^{-\frac{1}{2} + \frac{\beta_g}{1+2\beta_g} \frac{1}{\beta} \beta(1-4\beta/d) \frac{1+2\beta_g}{2\beta_g}} = \log(n) n^{-2\beta/d}$ , which agrees (up to a log factor) with the rate of  $n^{-2\beta/d}$  given by the simple estimator of Wang et al. (2006) analyzed

above under "Example 1b (cont)".

### Improved Rates of Convergence with $X$ random in a semiparametric

**model:** We now, as promised in the Introduction, construct an estimator of  $\sigma^2$  under the homoscedastic model  $E[Y|X] = b(X)$ ,  $\text{var}[Y|X] = \sigma^2$  with  $X$  random with unknown density that, whenever  $\beta < \min\{1, d/4\}$  and, regardless of the smoothness of  $f_X(x)$ , converges at the rate  $n^{-\frac{4\beta/d}{4\beta/d+1}}$ , which is faster than equal-spaced non-random minimax rate of  $n^{-2\beta/d}$ . Specifically we divide the support of  $X$ , i.e., the unit cube in  $R^d$ , into  $k = k(n) = n^\gamma$ ,  $\gamma > 1$  identical subcubes with edge length  $k^{-1/d}$ . We continue to assume the unknown density  $f_X(x)$  is absolutely continuous wrt to Lebesgue measure and both it and its inverse are bounded in sup-norm. Then it is a standard probability calculation that the number of subcubes containing at least two observations is  $O_p(n^2/k)$ . We estimate  $\sigma^2$  in each such subcube by  $(Y_i - Y_j)^2/2$ , where, for any subcube with 3 or more observations,  $i$  and  $j$  are chosen randomly, without replacement. Our final estimator of  $\sigma^2$  is the average of our subcube-specific estimates  $(Y_i - Y_j)^2/2$  over the  $O_p(n^2/k)$  subcubes with at least two observations. The rate of convergence of the estimator is minimized at  $n^{-\frac{4\beta/d}{4\beta/d+1}}$  by taking  $k = n^{\frac{2}{1+4\beta/d}}$ , as we now show.

We note that  $E[(Y_i - Y_j)^2/2 | X_i, X_j] = \sigma^2 + \{b(X_i) - b(X_j)\}^2/2$ ,  $|b(X_i) - b(X_j)| = O\|X_i - X_j\|^\beta$  by  $\beta < 1$ , and  $\|X_i - X_j\| = d^{1/2}O(k^{-1/d})$  when  $X_i$  and  $X_j$  are in the same subcube. It follows that the estimator has variance  $O_p(k/n^2)$  and bias of

$O(k^{-2\beta/d})$ . To minimize the convergence rate we equate the orders of the variance and the squared bias by solving  $k/n^2 = k^{-4\beta/d}$  which gives  $k = n^{\frac{2}{1+4\beta/d}}$ . Our random design estimator has better bias control and hence converges faster than the optimal equal-spaced fixed  $X$  estimator, because the random design estimator exploits the  $O_p\left(n^2/n^{\frac{2}{1+4\beta/d}}\right)$  random fluctuations for which  $X$ 's corresponding to two different observations are a distance of  $O\left(\left\{n^{\frac{2}{1+4\beta/d}}\right\}^{-1/d}\right)$  apart. Our estimator will not converge at rate  $n^{-\frac{4\beta/d}{4\beta/d+1}}$  to  $E[\text{var}(Y|X)]$  in our nonparametric model, because it then no longer suffices to average estimates of  $\text{var}(Y|X)$  only over subcubes containing 2 or more observations.

## 4.1 More Efficient Estimators

### 4.1.1 Case 1: The estimation bias of the third order estimator is less than the optimal rate

In a (locally) nonparametric model  $\mathcal{M}(\Theta)$ , the estimator  $\hat{\psi}_{m,k} = \hat{\psi} + \hat{\mathbb{I}\mathbb{F}}_{m,\tilde{\psi}_k}$  is essentially the unique  $m$ -th order U-statistic estimator of the truncated parameter  $\tilde{\psi}_k$  for which the leading term in the bias is  $O\left(\left\|\hat{\theta} - \theta\right\|^{m+1}\right)$ . However, when the minimax rate of convergence for  $\psi$  is slower than  $n^{-1/2}$ , other  $m^{\text{th}}$  order U-statistics estimators will often converge to  $\tilde{\psi}_k$  (and thus  $\psi$ ) at a faster rate uniformly over the model than does any estimator  $\hat{\psi}_{m,k}$  (constructed from an estimated higher order influence function  $\hat{\mathbb{I}\mathbb{F}}_{m,\tilde{\psi}_k}$  for  $\tilde{\psi}_k$ ) by tolerating bias at orders less than  $m+1$  in exchange for a

savings in variance.

**Remark 30** *A heuristic understanding as to why this is so can be gained from the following considerations. The theory of higher order influence functions as developed in theorems 2.2 and 2.3 is a theory of score functions (derivatives). Thus it can directly incorporate the restriction that a function, say  $b(x)$ , has an expansion  $b(x) = \sum_{l=1}^{\infty} \eta_l z_l(x)$  for which  $\eta_l = 0$  for  $l > k$ , as the restriction is equivalent to various scores being equal to zero. However the theory cannot directly incorporate restrictions such as  $\sum_{l=k}^{\infty} \eta_l^2 = k^{-2\beta_b}$  or  $\eta_l \propto l^{-(\beta_b + \frac{1}{2})}$  that do not imply any restrictions on score functions. Thus to find an optimal estimator, one must perform additional “side calculations” to quantify the estimation and truncation bias of various candidate estimators under these restrictions. As the assumption that  $b(x)$  lies in a Holder ball can be expressed in terms of such restrictions, this remark is relevant to a search for an optimal rate estimator.*

We now construct such estimators. We first consider the case where  $\beta_b, \beta_b$ , and  $\beta_g$  are such that the estimation bias  $O\left(n^{-\left(\frac{\beta_g}{2\beta_g+d} + \frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p}\right)}\right)$  of the second order estimator is greater than  $O\left(n^{-\frac{4\beta}{4\beta+d}}\right)$  but the estimation bias  $O\left(n^{-\left(\frac{2\beta_g}{2\beta_g+d} + \frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p}\right)}\right)$  of the third order estimator is less than  $O\left(n^{-\frac{4\beta}{4\beta+d}}\right)$ . That is

$$n^{-\left(\frac{2\beta_g}{2\beta_g+d} + \frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p}\right)} < n^{-\frac{4\beta}{4\beta+d}} < n^{-\left(\frac{\beta_g}{2\beta_g+d} + \frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p}\right)} \quad (40)$$

Then the most efficient estimator  $\widehat{\psi}_{m,k}$  in our class has rate of convergence slower than  $n^{-\frac{4\beta}{4\beta+d}}$  because  $\widehat{\psi}_{2,k_{opt}(2)}$  converges at rate  $n^{-\left(\frac{\beta g}{2\beta g+d} + \frac{\beta b}{d+2\beta b} + \frac{\beta p}{d+2\beta p}\right)}$  determined by the 2nd order estimation bias and, for  $m > 3$ ,  $\widehat{\psi}_{m,k_{opt}(m)}$  converges at a rate no faster than  $n^{-\frac{6\beta}{(d+2\beta)}} = n^{-4\frac{\beta}{d}3/((3-1)+4\frac{\beta}{d})} = \min_{\{m;m>3\}} n^{-4\frac{\beta}{d}m/((m-1)+4\frac{\beta}{d})}$ . [We obtained  $n^{-4\frac{\beta}{d}m/((m-1)+4\frac{\beta}{d})}$  as  $(k^{-4\beta/d})^{1/2}$ , where  $k$  solves the equation  $k^m/n^{m+1} = k^{-4\beta/d}$  that equates the variance  $k^m/n^{m+1}$  of  $\mathbb{IF}_m$  to the squared truncation bias  $k^{-4\beta/d}$ .]

To describe our more efficient estimator, define for nonnegative integers  $k(0), k(1), k^*(0), k^*(1)$  with  $k(0) < k(1)$  and  $k^*(0) < k^*(1)$  the  $U$ -statistic

$$\widehat{U}_3 \left( \begin{smallmatrix} k(1), k^*(1) \\ k(0), k^*(0) \end{smallmatrix} \right) = \mathbb{V} \left( \widehat{U}_3 \left( \begin{smallmatrix} k(1), k^*(1) \\ k(0), k^*(0) \end{smallmatrix} \right) \right)$$

with

$$\begin{aligned} & \widehat{U}_3 \left( \begin{smallmatrix} k(1), k^*(1) \\ k(0), k^*(0) \end{smallmatrix} \right) \\ &= \widehat{\epsilon}_{i_1} \overline{Z}_{k(0), i_1}^{k(1), T} \left( \left[ \dot{P} \dot{B} H_1 \overline{Z}_{k(0)}^{k(1)} \overline{Z}_{k^*(0)}^{k^*(1), T} \right]_{i_2} - I_{\{k(1)-k(0)\} \times \{k^*(1)-k^*(0)\}} \right) \overline{Z}_{k^*(0), i_3}^{k^*(1)} \widehat{\Delta}_{i_3} \\ &= \sum_{s_1=k(0)+1}^{k(1)} \sum_{s_2=k^*(0)+1}^{k^*(1)} \left\{ \widehat{\epsilon}_{i_1} z_{s_1}(X_{i_1}) \times \left\{ \left[ \dot{B} \dot{P} H_1 \right]_{i_2} z_{s_1}(X_{i_2}) z_{s_2}(X_{i_2}) - I[s_1 = s_2] \right\} z_{s_2}(X_{i_3}) \widehat{\Delta}_{i_3} \right\}, \end{aligned}$$

where  $\overline{Z}_{k(0)}^{k(1)} = (Z_{k(0)+1}, \dots, Z_{k(1)})^T$ ,  $\widehat{\epsilon} = (H_1 \widehat{P} + H_2) \dot{B}$ ,  $\widehat{\Delta} = (H_1 \widehat{B} + H_3) \dot{P}$ ,

$I_{r \times v} = (I_{ij})_{r \times v}$  with  $I_{ij} = I(i = j)$ .

As an example  $\widehat{\mathbb{IF}}_{33, \widetilde{\psi}_k} = \widehat{U}_3 \left( \begin{smallmatrix} k(1), k^*(1) \\ k(0), k^*(0) \end{smallmatrix} \right)$ . We can identify  $\left( \begin{smallmatrix} k(1), k^*(1) \\ k(0), k^*(0) \end{smallmatrix} \right)$  with the rectangle in  $R^2$  defined by  $\{(r_1, r_2); k(0) + 1 \leq r_1 \leq k(1), k^*(0) + 1 \leq r_1 \leq k^*(1)\}$  with

$(k(0) + 1, k^*(0) + 1)$  and  $(k(1) + 1, k^*(1) + 1)$ , respectively, the vertices closest and furthest from the origin. Thus  $\widehat{\mathbb{I}\mathbb{F}}_{33, \tilde{\psi}_k} = \widehat{\mathbb{U}}_3 \binom{k}{0}, \binom{k}{0}$  is identified with the rectangle  $\binom{k}{0}, \binom{k}{0}$ . Indeed we can write

$$\begin{aligned} & \widehat{\mathbb{U}}_3 \binom{k(1), k^*(1)}{k(0), k^*(0)} \\ &= \sum_{(s_1, s_2) \in \binom{k(1), k^*(1)}{k(0), k^*(0)}} \left\{ \widehat{\epsilon}_{i_1} z_{s_1}(X_{i_1}) \times \right. \\ & \quad \left. \left\{ \left[ \dot{B} \dot{P} H_1 \right]_{i_2} z_{s_1}(X_{i_2}) z_{s_2}(X_{i_2}) - I[s_1 = s_2] \right\} z_{s_2}(X_{i_3}) \widehat{\Delta}_{i_3} \right\} \end{aligned}$$

where, here and below,  $s_1$  and  $s_2$  are restricted to be integers, so  $(s_1, s_2) \in \binom{k(1), k^*(1)}{k(0), k^*(0)}$  are the lattice points of the rectangle.

We next study the variance of  $\widehat{\mathbb{U}}_3 \binom{k(1), k^*(1)}{k(0), k^*(0)}$ . It follows from Theorem 26 above that the number of lattice points in  $\binom{k(1), k^*(1)}{k(0), k^*(0)}$  is proportional to the variance of  $\widehat{\mathbb{U}}_3 \binom{k(1), k^*(1)}{k(0), k^*(0)}$  so if  $k(0) \ll k(1)$  and  $k^*(0) \ll k^*(1)$  then  $\text{var} \left[ \widehat{\mathbb{U}}_3 \binom{k(1), k^*(1)}{k(0), k^*(0)} \right]$  and  $\text{var} \left[ \widehat{\mathbb{U}}_3 \binom{k(1), k^*(1)}{0, 0} \right]$  are both of order  $k(1)k^*(1)/n^3$ . Hence the order of the variance of  $\widehat{\mathbb{U}}_3 \binom{k(1), k^*(1)}{k(0), k^*(0)}$  is determined by the vertex of the rectangle  $\binom{k(1), k^*(1)}{k(0), k^*(0)}$  furthest from the origin.

In contrast by a theorem in the appendix, the mean  $E \left[ \widehat{\mathbb{U}}_3 \binom{k(1), k^*(1)}{k(0), k^*(0)} \right]$  is

$$\widehat{E} \left( \widehat{\Pi} \left[ \delta b | \overline{Z}_{k(0)}^{k(1)} \right] \delta g \widehat{Q}^2 \widehat{\Pi} \left[ \delta p | \overline{Z}_{k^*(0)}^{k^*(1)} \right] \right) (1 + o_p(1))$$

with  $\delta b = \dot{P} \widehat{E}(H_1|X) (\widehat{B} - B)$ ,  $\delta p = \dot{B} \widehat{E}(H_1|X) (\widehat{P} - P)$ ,  $\delta g = \frac{g(X) - \widehat{g}(X)}{\widehat{g}(X)}$  and  $\widehat{Q}^2 = \dot{B} \dot{P} \widehat{E}(H_1|X)$ . It follows that if  $k(0) \ll k(1)$  and  $k^*(0) \ll k^*(1)$  then



$E \left[ \widehat{\mathbb{U}}_3 \left( \begin{smallmatrix} k(1), k^*(1) \\ k(0), k^*(0) \end{smallmatrix} \right) \right]$  and  $E \left[ \widehat{\mathbb{U}}_3 \left( \begin{smallmatrix} \infty, \infty \\ k(0), k^*(0) \end{smallmatrix} \right) \right]$  are both of order

$$O_p \left[ k(0)^{-\beta_b} k^*(0)^{-\beta_p} (n/\log n)^{\frac{-\beta_g}{2\beta_g+1}} \right].$$

To see this for  $E \left[ \widehat{\mathbb{U}}_3 \left( \begin{smallmatrix} k(1), k^*(1) \\ k(0), k^*(0) \end{smallmatrix} \right) \right]$ , we ‘sup out’  $|\delta g \widehat{Q}^2|$  from  $\widehat{E} \left( \left| \widehat{\Pi} \left[ \delta b | \overline{Z}_{k(0)}^{k(1)} \right] \delta g \widehat{Q}^2 \widehat{\Pi} \left[ \delta p | \overline{Z}_{k^*(0)}^{k^*(1)} \right] \right| \right)$

which is

$$O_p \left[ (n/\log n)^{\frac{-\beta_g}{2\beta_g+1}} \right] \widehat{E} \left( \left| \widehat{\Pi} \left[ \delta b | \overline{Z}_{k(0)}^{k(1)} \right] \widehat{\Pi} \left[ \delta p | \overline{Z}_{k^*(0)}^{k^*(1)} \right] \right| \right).$$

We then apply Cauchy Schwartz to  $\widehat{E} \left( \left| \widehat{\Pi} \left[ \delta b | \overline{Z}_{k(0)}^{k(1)} \right] \widehat{\Pi} \left[ \delta p | \overline{Z}_{k^*(0)}^{k^*(1)} \right] \right| \right)$ , noting that  $\widehat{E} \left( \left\{ \widehat{\Pi} \left[ \delta b | \overline{Z}_{k(0)}^{k(1)} \right] \right\}^2 \right)^{1/2} = O \left( k(0)^{-\beta_b} \right)$ . Again a more careful argument using Hölder’s

inequality would show the log factor is unnecessary. Hence the order of the mean of

$\widehat{\mathbb{U}}_3 \left( \begin{smallmatrix} k(1), k^*(1) \\ k(0), k^*(0) \end{smallmatrix} \right)$  is determined by the vertex of the rectangle  $\left( \begin{smallmatrix} k(1), k^*(1) \\ k(0), k^*(0) \end{smallmatrix} \right)$  closest to the origin.

**Motivation:** With this background we are ready to motivate our new estimator. Recall from Section 3.2.5, that with  $g$  known, the choice  $k_{opt}^g(2) = n^{\frac{2}{1+4\beta/d}}$  gives  $\left( \widehat{\psi}_{2, k_{opt}^g(2)} - \psi \right) = O_p \left( n^{-\frac{4\beta}{4\beta+d}} \right)$  because the truncation bias  $\left| \widetilde{\psi}_{k_{opt}^g(2)} - \psi \right|$  and variance are of order  $n^{-\frac{4\beta}{4\beta+d}}$  and the estimation bias is zero. Any choice of  $k$  larger than  $k_{opt}^g(2)$  will result in a slower rate of convergence.

However, when  $g$  is unknown and thus estimated,  $\widehat{\psi}_{2, k_{opt}^g(2)} - \psi$  does not attain the optimal rate of convergence because the estimation bias  $n^{-\left( \frac{\beta_g}{2\beta_g+d} + \frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p} \right)}$  exceeds  $n^{-\frac{4\beta}{4\beta+d}}$ . The estimator  $\widehat{\psi}_{3, k_{opt}^g(2)} = \widehat{\psi}_{2, k_{opt}^g(2)} + \widehat{\mathbb{U}}_3 \left( \begin{smallmatrix} k_{opt}^g(2), k_{opt}^g(2) \\ 0, 0 \end{smallmatrix} \right)$  also fails to attain

the rate  $n^{-\frac{4\beta}{4\beta+d}}$  because it has variance of the order of

$$\frac{k_{opt}^g(2)}{n} \frac{k_{opt}^g(2)}{n^2} = O\left(\frac{n^{\frac{2}{1+4\beta/d}}}{n} n^{-\frac{8\beta}{4\beta+d}}\right),$$

which exceeds  $O\left(n^{-\frac{8\beta}{4\beta+d}}\right)$ . On the other hand,  $\widehat{\psi}_{3,k_{opt}^g(2)}$  has bias of  $O_p\left(n^{-\frac{4\beta}{4\beta+d}}\right)$  because the truncation bias is  $O_p\left(n^{-\frac{4\beta}{4\beta+d}}\right)$  and the estimation bias  $O_p\left(n^{-\left(\frac{2\beta_g}{2\beta_g+d} + \frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p}\right)}\right)$  is also  $O_p\left(n^{-\frac{4\beta}{4\beta+d}}\right)$  under our assumption (40). Our strategy will be to try to replace the term  $\widehat{\mathbb{U}}_3\left(\begin{smallmatrix} k_{opt}^g(2) \\ 0, \end{smallmatrix} k_{opt}^g(2) \right)$  in the estimator  $\widehat{\psi}_{3,k_{opt}^g(2)} = \widehat{\psi}_{2,k_{opt}^g(2)} + \widehat{\mathbb{U}}_3\left(\begin{smallmatrix} k_{opt}^g(2) \\ 0, \end{smallmatrix} k_{opt}^g(2) \right)$  by

$$\widehat{U}_3(\Omega) = \sum_{(s_1, s_2) \in \Omega} \left\{ \widehat{\epsilon}_{i_1} z_{s_1}(X_{i_1}) z_{s_2}(X_{i_3}) \widehat{\Delta}_{i_3} \times \left\{ \left[ \dot{B} \dot{P} H_1 \right]_{i_2} z_{s_1}(X_{i_2}) z_{s_2}(X_{i_2}) - I[s_1 = s_2] \right\} \right\}$$

where  $\Omega$  is a subset of the rectangle  $\left(\begin{smallmatrix} k_{opt}^g(2) \\ 0, \end{smallmatrix} k_{opt}^g(2) \right)$  such that  $var\left(\widehat{U}_3(\Omega)\right) \asymp n^{-\frac{8\beta}{4\beta+d}}$  but the additional bias

$$\begin{aligned} & E\left[\widehat{\mathbb{U}}_3\left(\begin{smallmatrix} k_{opt}^g(2) \\ 0, \end{smallmatrix} k_{opt}^g(2) \right) - \widehat{U}_3(\Omega)\right] \\ &= E\left[\widehat{\mathbb{U}}_3\left(\left(\begin{smallmatrix} k_{opt}^g(2) \\ 0, \end{smallmatrix} k_{opt}^g(2) \right) \setminus \Omega\right)\right] \\ &\equiv E\left[\sum_{(s_1, s_2) \in \left(\begin{smallmatrix} k_{opt}^g(2) \\ 0, \end{smallmatrix} k_{opt}^g(2) \right) \setminus \Omega} \left\{ \widehat{\epsilon}_{i_1} z_{s_1}(X_{i_1}) \left\{ \begin{aligned} & \left[ \dot{B} \dot{P} H_1 \right]_{i_2} z_{s_1}(X_{i_2}) z_{s_2}(X_{i_2}) \\ & - I[s_1 = s_2] \end{aligned} \right\} \right\} \right. \\ & \quad \left. \times z_{s_2}(X_{i_3}) \widehat{\Delta}_{i_3} \right\} \right] \end{aligned}$$

is  $O_p\left(n^{-\frac{4\beta}{4\beta+d}}\right)$ . This approach will succeed if we can chose  $\Omega$  and thus  $\left(\begin{smallmatrix} k_{opt}^g(2) \\ 0, \end{smallmatrix} k_{opt}^g(2) \right) \setminus \Omega$

to be sums of rectangles (whose number does not increase with  $n$ ) such that (i) each

rectangle in  $\left(0, k_{opt}^g(2), k_{opt}^g(2)\right) \setminus \Omega$  has its closest vertex to the origin, say  $(k(0), k^*(0))$ , satisfying  $O_p \left[ k(0)^{-\beta_b} k^*(0)^{-\beta_p} n^{\frac{-\beta_g}{2\beta_g+1}} \right] \leq n^{-\frac{4\beta}{4\beta+d}}$  and (ii) simultaneously each rectangle in  $\Omega$  has its furthest vertex from the origin, say  $(k(1), k^*(1))$ , satisfying  $O(k(1)k^*(1)/n^3) = O\left(n^{-\frac{8\beta}{4\beta+d}}\right)$ .

We index the vertices of our set of rectangles as follows. Consider a natural number  $J$  and a set of non-negative integers  $\mathcal{K}_{J,tot} = \{k_{-2}, k_{-1}, k_0, \dots, k_{2J}, k_{2J+1}, k_{2J+2}\}$  satisfying  $0 = k_{-2} < k_0 < k_2 < \dots < k_{2J-2} < k_{2J} < k_{2J+2} = k_{2J+1} < k_{2J-1} < \dots < k_1 < k_{-1}$

Note the elements with even subscripts increase from 0 to  $2J+2$  while elements with odd subscripts decrease from  $-1$  to  $2J-1$ . Further the smallest element with odd subscript equals the largest element with even subscript. We will use two such sets  $\mathcal{K}_{b,J,tot}$  and  $\mathcal{K}_{p,J,tot}$  with corresponding elements  $k_{bl}$  and  $k_{pl}$  with  $k_{b,-1} = k_{p,-1}$ .

Set for  $s \in \{-1, 0, \dots, J\}$

$$k_{b,2s+1} = n^{\frac{3d+4\beta}{(d+4\beta)}} / k_{p,2s+2}, \quad (41)$$

$$k_{p,2s+1} = n^{\frac{3d+4\beta}{(d+4\beta)}} / k_{b,2s+2}, \text{ so} \quad (42)$$

$$\frac{k_{p,2s+1}k_{b,2s+2}}{n^3} = \frac{k_{b,2s+1}k_{p,2s+2}}{n^3} = n^{-\frac{8\beta}{4\beta+d}}$$

We leave  $J, \mathcal{K}_{p,J} = \{k_{p,2s}, s = 0, \dots, J+1\}$ , and  $\mathcal{K}_{b,J} = \{k_{b,2s}, s = 0, \dots, J+1\}$  unspecified for now but derive optimal values below.

Let  $\Omega = \Omega(\mathcal{K}_{p,J}, \mathcal{K}_{b,J})$  be the union of rectangles

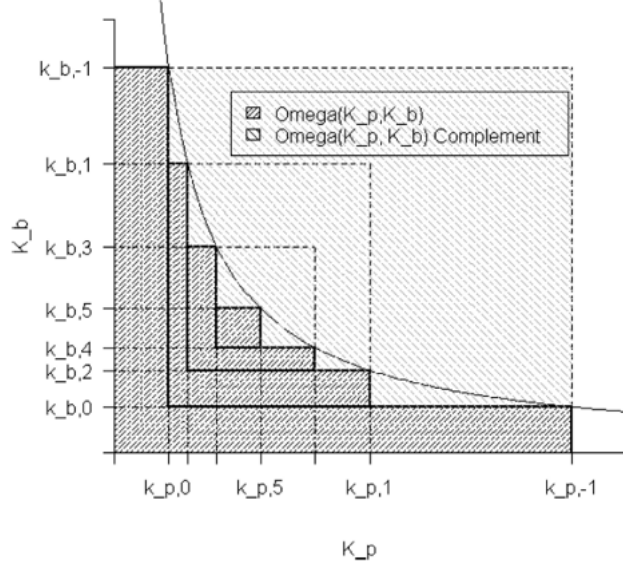


FIG 1. Hyperbola  $H_y$  and Associated Rectangles.

$$\Omega(\mathcal{K}_{pJ}, \mathcal{K}_{bJ}) = \left\{ \bigcup_{s=0}^J \begin{pmatrix} k_{p,2s-1}, k_{b,2s} \\ k_{p,2s-2}, k_{b,2s-2} \end{pmatrix} \cup \begin{pmatrix} k_{p,2s}, k_{b,2s-1} \\ k_{p,2s-2}, k_{b,2s} \end{pmatrix} \right\} \cup \begin{pmatrix} k_{p,2J+1}, k_{b,2J+1} \\ k_{p,2J}, k_{b,2J} \end{pmatrix}$$

The points  $(k_{p,2s+1}, k_{b,2s+2}), (k_{p,2s+2}, k_{b,2s+1})$  for  $s = -1, 0, \dots, J+1$  lie on a hyperbola  $H_y$  in  $R^2$  defined by  $H_y = \left\{ (r_1, r_2); r_1 r_2 = n^{\frac{3d+4\beta}{(d+4\beta)}} \right\}$  shown in figure 1 for  $J = 2$ . The set  $\Omega(\mathcal{K}_{pJ}, \mathcal{K}_{bJ}) \subset \left( 0, k_{opt}^g(2) \right) \times \left( 0, k_{opt}^g(2) \right)$  lies below  $H_y$ .

Define

$$\widehat{\psi}_{3,(\mathcal{K}_{pJ}, \mathcal{K}_{bJ})} = \widehat{\psi}_{2, k_{-1}} + \widehat{\mathbb{U}}_3(\Omega(\mathcal{K}_{pJ}, \mathcal{K}_{bJ})).$$

We then have

**Theorem 31** (i): The estimator  $\widehat{\psi}_{3,(\mathcal{K}_{pJ}, \mathcal{K}_{bJ})}$  has variance of the order of

$$\frac{k_{-1}}{n^2} + (2J+1)n^{-\frac{8\beta}{4\beta+d}}$$

and bias  $E\left(\widehat{\psi}_{3,(\mathcal{K}_{pJ}, \mathcal{K}_{bJ})}\right) - \psi$  of order

$$\begin{aligned} & O_p \left\{ n^{-\frac{\beta_g}{2\beta_g+d}} \left( \sum_{s=0}^J \left( k_{p,2s+1}^{-\beta_p/d} k_{b,2s}^{-\beta_b/d} + k_{b,2s+1}^{-\beta_b/d} k_{p,2s}^{-\beta_p/d} \right) \right) \right\} \\ & + O_p \left( n^{-\left(\frac{2\beta_g}{2\beta_g+d} + \frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p}\right)} \right) + O_p \left( k_{-1}^{-(\beta_p+\beta_b)/d} \right) \end{aligned}$$

Proof: Each of the  $2J+1$  rectangles whose union is  $\Omega(\mathcal{K}_{pJ}, \mathcal{K}_{bJ})$  has  $(k_{p,2s+1}, k_{b,2s+2})$  or  $(k_{p,2s+2}, k_{b,2s+1})$  for some  $s \in \{-1, 0, \dots, J\}$  as the vertex furthest from the origin and thus contributes  $\frac{k_{p,2s+1}k_{b,2s+2}}{n^3} = n^{-\frac{8\beta}{4\beta+d}}$  to the variance of  $\widehat{\psi}_{3,(\mathcal{K}_{pJ}, \mathcal{K}_{bJ})}$ . The variance of  $\widehat{\psi}_{2,k_{-1}} \asymp \frac{k_{-1}}{n^2}$ . Now

$$\begin{aligned} & E\left(\widehat{\psi}_{3,(\mathcal{K}_{pJ}, \mathcal{K}_{bJ})}\right) - \psi \\ & = \left\{ E\left(\widehat{\psi}_{3,k_{-1}}\right) - \psi \right\} + \left\{ E\left[\widehat{\mathbb{U}}_3\left\{\Omega((\mathcal{K}_{pJ}, \mathcal{K}_{bJ}))\right\}\right] - E\left[\widehat{\mathbb{U}}_3\left\{\left(\begin{smallmatrix} k_{-1} & k_{-1} \\ 0 & 0 \end{smallmatrix}\right)\right\}\right] \right\} \\ & = O_p\left(k_{-1}^{-(\beta_p+\beta_b)/d}\right) + O_p\left(n^{-\left(\frac{2\beta_g}{2\beta_g+d} + \frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p}\right)}\right) \\ & + E\left[\widehat{\mathbb{U}}_3\left\{\left(\begin{smallmatrix} k_{-1} & k_{-1} \\ 0 & 0 \end{smallmatrix}\right) \setminus \Omega((\mathcal{K}_{pJ}, \mathcal{K}_{bJ}))\right\}\right] \end{aligned}$$

As is evident from Figure 1,  $\Omega^c(\mathcal{K}_{pJ}, \mathcal{K}_{bJ}) \equiv \left(\begin{smallmatrix} k_{-1} & k_{-1} \\ 0 & 0 \end{smallmatrix}\right) \setminus \Omega((\mathcal{K}_{pJ}, \mathcal{K}_{bJ}))$  is the union of rectangles  $\cup_{s=0}^J \left\{ \left(\begin{smallmatrix} k_{p,2s-1} & k_{b,2s-1} \\ k_{p,2s} & k_{b,2s+1} \end{smallmatrix}\right) \cup \left(\begin{smallmatrix} k_{p,2s-1} & k_{b,2s+1} \\ k_{p,2s+1} & k_{b,2s} \end{smallmatrix}\right) \right\}$  which have

$$\{(k_{p,2s}, k_{b,2s+1}), (k_{p,2s+1}, k_{b,2s}); s \in \{-1, 0, \dots, J\}\}$$

as the set of vertices closest to the origin, leading to the expression for the bias given in the theorem.

**Theorem 32** *Given  $(\beta_b, \beta_p, \beta_g)$  with  $\beta_p \geq \beta_b$  so  $\Delta \geq 0$ , Eq.(37) holds if and only if there exists  $J, \mathcal{K}_{pJ}, \mathcal{K}_{bJ}$  such that  $\widehat{\psi}_{3,(\mathcal{K}_{pJ}, \mathcal{K}_{bJ})} - \psi = O_p\left(n^{-\frac{4\beta}{4\beta+d}}\right)$ .*

*If Eq.(37) holds,  $E\left[\widehat{\mathbb{U}}_3\left\{\binom{k_{-1}, k_{-1}}{0, 0} \setminus \Omega((\mathcal{K}_{pJ}, \mathcal{K}_{bJ}))\right\}\right] = O_p\left(n^{-\frac{4\beta}{4\beta+d}}\right)$  and thus  $\widehat{\psi}_{3,(\mathcal{K}_{pJ}, \mathcal{K}_{bJ})} - \psi = O_p\left(n^{-\frac{4\beta}{4\beta+d}}\right)$ , when we choose  $J$  to be the smallest integer such that*

$$(1 + \Delta)(J + 1) + c^*(\beta_g, \beta, \Delta) \sum_{l=1}^{J+1} (1 + \Delta)^{l-1} > \frac{3+4\beta/d}{2(1+4\beta/d)} \text{ with}$$

$$c^*(\beta_g, \beta, \Delta) = \left(\frac{2\beta_g/d}{2\beta_g/d + 1}\right) \frac{(\Delta + 2)}{4\beta/d} - \frac{2(\Delta + 2)}{4\beta/d + 1} + \frac{3 + 4\beta/d}{(1 + 4\beta/d)},$$

$$k_{b,0} = k_{p,0} = n, \quad k_{b,2s} = k_{p,2s} = n^{(1+\Delta)s} n^{q \sum_{l=1}^s (1+\Delta)^{l-1}} \text{ for } s = 1, \dots, J+1, \text{ with}$$

$$q = \left\{ \frac{3+4\beta/d}{2(1+4\beta/d)} - (1 + \Delta)(J + 1) \right\} / \sum_{l=1}^{J+1} (1 + \Delta)^{l-1}.$$

*Note  $J$  does not depend on the sample size  $n$ .*

Proof: From Theorem 31, for the variance of  $\widehat{\psi}_{3,(\mathcal{K}_{pJ}, \mathcal{K}_{bJ})}$  to be  $O_p\left(n^{-\frac{8\beta}{4\beta+d}}\right)$ ,  $J$  cannot increase with  $n$ . Further for the second order truncation bias  $O_p\left(k_{-1}^{-(\beta_p + \beta_b)/d}\right)$  and the square root of the variance  $\frac{k_{-1}}{n^2}$  of  $\widehat{\psi}_{2,k_{-1}}$  both to be  $O_p\left(n^{-\frac{4\beta}{4\beta+d}}\right)$ , we must have  $k_{-1} = k_{opt}^g(2) = n^{\frac{2}{1+4\beta/d}}$ . It then follows from eqs. (41) and (42) that  $k_{p,0} = k_{b,0} = n$ .

In order for  $E\left[\widehat{\mathbb{U}}_3\left\{\binom{k_{-1}, k_{-1}}{0, 0} \setminus \Omega((\mathcal{K}_{pJ}, \mathcal{K}_{bJ}))\right\}\right] = O_p\left(n^{-\frac{4\beta}{4\beta+d}}\right)$ , we require for  $s = 0, \dots, J$

$$n^{-\frac{2\beta_g}{2\beta_g+d}} \left\{ k_{b,2s}^{-2\beta_b/d} k_{p,2s+1}^{-2\beta_p/d} \right\} \leq n^{-\frac{8\beta/d}{4\beta/d+1}} \quad (43)$$

$$n^{-\frac{2\beta_g}{2\beta_g+d}} \left\{ k_{p,2s}^{-2\beta_p/d} k_{b,2s+1}^{-2\beta_b/d} \right\} \leq n^{-\frac{8\beta/d}{4\beta/d+1}} \quad (44)$$

Substituting for  $k_{b,2s+1}$  in eq. (44) using eq.(41) and recalling that  $\beta_p \geq \beta_b$  so  $\Delta \geq 0$ ,

we obtain

$$\begin{aligned} & n^{-\frac{2\beta_g}{2\beta_g+d}} k_{p,2s}^{-2\beta_p/d} \left\{ \frac{n^{\frac{3d+4\beta}{(d+4\beta)}}}{k_{p,2s+2}} \right\}^{-2\beta_b/d} \leq n^{-\frac{8\beta}{4\beta+d}} \quad (45) \\ \Leftrightarrow & k_{p,2s+2}^{2\beta_b/d} \leq n^{\frac{2\beta_g/d}{2\beta_g/d+1}} n^{-\frac{8\beta/d}{4\beta/d+1}} k_{p,2s}^{2\beta_p/d} \left( n^{\frac{3+4\beta/d}{(1+4\beta/d)}} \right)^{2\beta_b/d} \\ \Leftrightarrow & k_{p,2s+2} \leq n^{\left(\frac{2\beta_g/d}{2\beta_g/d+1}\right) \frac{1}{2\beta_b/d}} n^{-\frac{8\beta/d}{4\beta/d+1} \frac{1}{2\beta_b/d}} k_{p,2s}^{\frac{\beta_p}{\beta_b}} \left( n^{\frac{3+4\beta/d}{(1+4\beta/d)}} \right) \\ \Leftrightarrow & 1 \leq \frac{k_{p,2s+2}}{k_{p,2s}} \leq n^{\left(\frac{2\beta_g/d}{2\beta_g/d+1}\right) \frac{1}{2\beta_b/d}} n^{-\frac{8\beta/d}{4\beta/d+1} \frac{1}{2\beta_b/d}} k_{p,2s}^{\Delta} \left( n^{\frac{3+4\beta/d}{(1+4\beta/d)}} \right) \\ \Leftrightarrow & 1 \leq \frac{k_{p,2s+2}}{k_{p,2s}} \leq n^{c^*(\beta_g, \beta, \Delta)} k_{p,2s}^{\Delta} \quad (46) \\ \Leftrightarrow & 1 \leq n^{c^*(\beta_g, \beta, \Delta)} n^{\Delta} \\ \Leftrightarrow & 0 \leq c^*(\beta_g, \beta, \Delta) + \Delta \end{aligned}$$

since  $n = k_0 \leq k_{p,2s} \leq k_{p,2s+2}$ .

Solving the last expression for  $\frac{2\beta_g/d}{2\beta_g/d+1}$ , we obtain

$$\frac{2\beta_g/d}{2\beta_g/d+1} \geq \frac{\frac{1-4\beta/d}{1+4\beta/d} + \Delta \left\{ \frac{2}{4\beta/d+1} - 1 \right\}}{\frac{(\Delta+2)}{4\beta/d}} = \left\{ \frac{4\beta/d}{(\Delta+2)} \right\} (\Delta+1) \frac{1-4\beta/d}{1+4\beta/d} \quad (47)$$

which is equation eq.(37), except with a nonstrict inequality. We have just deduced that the constraint (47) was due to restriction (44). We have not yet considered whether the restriction (43) implies additional constraints. We now show that it does not. Specifically if we set  $k_{p,2l} = k_{b,2l}$  for all  $l \in \{1, 2, \dots, J+1\}$ , then eq.(43) is true whenever eq.(44) holds because of our assumption that  $\Delta \geq 0$ . Thus we can set  $\mathcal{K}_{pJ} = \mathcal{K}_{bJ}$ .

Thus we have shown that if  $\widehat{\psi}_{3,(\mathcal{K}_{pJ}, \mathcal{K}_{bJ})} - \psi = O_p\left(n^{-\frac{4\beta}{4\beta+d}}\right)$ , then  $k_{-1} = n^{\frac{2}{1+4\beta/d}}$ , (47) holds, and  $J$  must not increase with  $n$ .

We next show that when the inequality is strict in (47) and eq.(40) holds, we can find  $\mathcal{K}_J = \mathcal{K}_{pJ} = \mathcal{K}_{bJ}$  for which  $\widehat{\psi}_{3,\mathcal{K}_J} - \psi = O_p\left(n^{-\frac{4\beta}{4\beta+d}}\right)$ . We then complete the proof of the theorem by showing that when (47) holds with an equality, there is no choice of  $\mathcal{K}_J$  for which  $\widehat{\psi}_{3,\mathcal{K}_J}$  converges at a rate better than  $O_p\left((\log n) n^{-\frac{4\beta}{4\beta+d}}\right)$ .

Suppose the inequality is strict in (47). Since  $k_0 = n$ , eq.(46) applied recursively suggests we define  $k_{2s} = n^{(1+\Delta)s} n^{c^*(\beta_g, \beta, \Delta) \sum_{l=1}^s (1+\Delta)^{l-1}}$  for  $s = 1, \dots, J+1$  and take  $k_{2s+1} = \frac{n^{\frac{3d+4\beta}{(d+4\beta)}}}{k_{2s+2}}$ . However, this will not generally give  $k_{2J+1} = k_{2J+2} = n^{\left\{\frac{3d+4\beta}{(d+4\beta)}\right\} \frac{1}{2}}$  as required when  $\mathcal{K}_{pJ} = \mathcal{K}_{bJ}$ . Instead we use the modified algorithm given in the statement of the theorem which insures that  $k_{2J+1} = k_{2J+2} = n^{\frac{3+4\beta/d}{2(1+4\beta/d)}}$ , as required. Since  $J$  is not a function of  $n$ , in order to show  $\widehat{\psi}_{3,\mathcal{K}_J}$  converges at rate  $n^{-\frac{4\beta}{4\beta+d}}$ , we only need to check the bias.



Now  $\frac{k_{2s+2}}{k_{2s}} = n^{(1+\Delta)} n^{q(1+\Delta)^{s-1}} = k_0^{(1+\Delta)} n^{q(1+\Delta)^{s-1}} \leq k_0^{(1+\Delta)} n^{c^*(\beta_g, \beta, \Delta)(1+\Delta)^{s-1}}$  since  $q \leq c^*(\beta_g, \beta, \Delta)$  so the bias of  $\widehat{\psi}_{3, \mathcal{K}_J}$  is  $O_P\left(n^{-\frac{4\beta}{4\beta+d}}\right)$ , as required.

Suppose now the equality holds in eq.(47) so  $c^*(\beta_g, \beta, \Delta) + \Delta = 0$  and continue to assume eq.(40) holds. We now construct an estimator  $\widehat{\psi}_{3, \mathcal{K}_J}$  that converges at rate  $O_P\left(n^{-\frac{4\beta}{4\beta+d}} \ln(n)\right)$  and show that no estimator in our class  $\widehat{\psi}_{3, \mathcal{K}_J}$  converges at a faster rate. We conjecture this rate is minimax when the equality in eq.(47) holds. Again  $k_{2s+1} = \frac{n^{\frac{3d+4\beta}{(d+4\beta)}}}{k_{2s+2}}$  and by the previous arguments,  $k_0 = n, k_{-1} = n^{\frac{2}{(1+4\beta/d)}}$ ,  $k_{2J+1} = k_{2J+2} = \left\{n^{\frac{3d+4\beta}{(d+4\beta)}}\right\}^{1/2}$ . We can suppose that  $k_{2s} = n\{v(n)\}^s$ . It remains to determine  $v(n)$  and  $J = J(n)$ . We know  $J(n)$  must satisfy

$$k_{2J(n)+2} = \left\{n^{\frac{3d+4\beta}{(d+4\beta)}}\right\}^{1/2} = n\{v(n)\}^{J(n)+1} \text{ so}$$

$$v(n) = n^{\left(\frac{3d+4\beta}{2(d+4\beta)} - 1\right) \frac{1}{J(n)+1}}.$$

The variance of  $\widehat{\psi}_{3, \mathcal{K}_J}$  is of order  $n^{-\frac{8\beta}{4\beta+d} J(n)}$ . Thus the order of the bias will still equal that of the variance provided we multiply the RHS of eq.(45) by  $J(n)$ . Then eq.(46) becomes  $1 \leq \frac{k_{p, 2s+2}}{k_{p, 2s}} \leq n^{c^*(\beta_g, \beta, \Delta)} k_{p, 2s}^\Delta J(n)^{\frac{1}{2\beta/d}}$ . Since,  $\frac{k_{p, 2s+2}}{k_{p, 2s}} = v(n)$  and  $n = k_0 \leq k_{p, 2s}$ , we substitute  $n^\Delta = k_0^\Delta$  for  $k_{p, 2s}^\Delta$  in the modified eq.(46) which gives  $v(n) = J(n)^{\frac{1}{2\beta/d}}$ . Hence  $n^{\left(\frac{3d+4\beta}{2(d+4\beta)} - 1\right) \frac{1}{J(n)+1}} = J(n)^{\frac{1}{2\beta/d}}$  which implies that.

$$\frac{\ln(n)}{J(n)} = O(\ln[J(n)]) \quad (48)$$

To minimize the variance, we want the slowest growing function of  $n$  that satisfies eq.(48), which is  $J(n) = \ln(n)$ , as claimed.

#### 4.1.2 Case 2: The estimation bias of the third order estimator exceeds the optimal rate

In this section we no longer assume that the estimation bias  $n^{-\left(\frac{2\beta_g}{2\beta_g+d} + \frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p}\right)}$  of a third order estimator is less than  $n^{-\frac{4\beta}{4\beta+d}}$ . Then even when eq.(47) holds with a strict inequality,  $\widehat{\psi}_{3,\mathcal{K}_J}$  does not achieve a  $n^{-\frac{4\beta}{4\beta+d}}$  rate of convergence because the fourth order bias  $n^{-\left(\frac{2\beta_g}{2\beta_g+d} + \frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p}\right)}$  exceeds  $n^{-\frac{4\beta}{4\beta+d}}$ . However, we will now construct an estimator  $\widehat{\psi}_{\mathcal{K}_J}^{eff} \equiv \widehat{\psi}_{\mathcal{K}_J}^{eff}(\beta_g, \beta_b, \beta_p)$  that under our assumptions  $Ai) - Aiv)$  does converge at rate  $n^{-\frac{4\beta}{4\beta+d}}$  whenever  $(\beta_g, \beta_b, \beta_p)$  given in assumption  $Aiv)$  satisfy eq.(47) with a strict inequality. Because the estimator is very complicated, we have chosen to only define the estimator and give its properties in the text. The motivating ideas for and the formal proofs of these properties are provided in the appendix.

To define the estimator, we need some additional notation. Define

$$\begin{aligned} & \widehat{\mathbb{U}}_m \left( (l)_{k(l,0)}^{k(l,1)}, 1 \leq l \leq m-1 \right) \\ &= \mathbb{V}_m \left( \widehat{\epsilon}_{i_1} \overline{Z}_{k(1,0),i_1}^{k(1,1)T} \prod_{u=2}^{m-1} \left( \dot{B} \dot{P} H_1 \overline{Z}_{k(u-1,0)}^{k(u-1,1)} \overline{Z}_{k(u,0)}^{k(u,1)T} - I_{k_{u-1} \times k_u} \right) \overline{Z}_{k(m-1,0)}^{k(m-1,1)} \widehat{\Delta}_{i_m} \right) \end{aligned}$$

where ,  $k_u = k(u, 1) - k(u, 0)$ ,  $I_{k_{u-1} \times k_u} = (I_{ij})_{k_{u-1} \times k_u}$  with  $I_{ij} = I(i = j)$ .

Then define  $\widehat{\mathbb{U}}_m \left( \begin{smallmatrix} k(1) \\ k(0) \end{smallmatrix} \right)$  as  $\widehat{\mathbb{U}}_m \left( (l)_{k(0)}^{k(1)}, 1 \leq l \leq m-1 \right)$ .  $\widehat{\mathbb{U}}_m^{(u)} \left( \begin{smallmatrix} k^*(1) & k(1) \\ k^*(0) & k(0) \end{smallmatrix} \right)$  is defined as  $\widehat{\mathbb{U}}_m \left( (l)_{k(l,0)}^{k(l,1)}, 1 \leq l \leq m-1 \right)$  with  $k(l, 1) = k(1)$ ,  $k(l, 0) = k(0)$  for  $l \neq u$ , and  $k(u, 1) = k^*(1)$ ,  $k(u, 0) = k^*(0)$ . Next  $\widehat{\mathbb{U}}_m^{(u,u+1)} \left( \begin{smallmatrix} k^*(1) & k^{**}(1) & k(1) \\ k^*(0) & k^{**}(0) & k(0) \end{smallmatrix} \right)$  is defined as  $\widehat{\mathbb{U}}_m \left( (l)_{k(l,0)}^{k(l,1)}, 1 \leq l \leq m-1 \right)$  with  $k(l, 1) = k(1)$ ,  $k(l, 0) = k(0)$  for  $l \neq u$  and  $l \neq u+1$ ,  $k(u, 1) = k^*(1)$ ,

$k(u, 0) = k^*(0)$ ,  $k(u+1, 1) = k^{**}(1)$ ,  $k(u+1, 0) = k^{**}(0)$ . We will use this notation for  $m = 3$ , even though  $\widehat{\mathbb{U}}_3^{(1,2)} \begin{pmatrix} k^*(1) & k^{**}(1) & k(1) \\ k^*(0) & k^{**}(0) & k(0) \end{pmatrix}$  does not depend on  $k(0), k(1)$  and is equal to  $\widehat{\mathbb{U}}_3 \begin{pmatrix} k^*(1) & k^{**}(1) \\ k^*(0) & k^{**}(0) \end{pmatrix}$  of the previous subsection.

Finally define

$$\begin{aligned} \mathbb{H}_v^* &= \widehat{\mathbb{U}}_v \begin{pmatrix} k_0 \\ 0 \end{pmatrix} + \sum_{u=1}^{v-1} \widehat{\mathbb{U}}_v^{(u)} \begin{pmatrix} k_{-1} & k_0 \\ k_0 & 0 \end{pmatrix} \\ \mathbb{G}(s, v) &= \sum_{u=1}^{v-2} \left\{ \widehat{\mathbb{U}}_v^{(u, u+1)} \begin{pmatrix} k_{2s-1} & k_{2s} & k_0 \\ k_{2s-2} & k_{2s-2} & 0 \end{pmatrix} + \widehat{\mathbb{U}}_v^{(u, u+1)} \begin{pmatrix} k_{2s} & k_{2s-1} & k_0 \\ k_{2s-2} & k_{2s} & 0 \end{pmatrix} \right\} \\ \mathbb{Q}_v &= \sum_{u=1}^{v-2} \widehat{\mathbb{U}}_v^{(u, u+1)} \begin{pmatrix} k_{2J+1} & k_{2J+1} & k_0 \\ k_{2J} & k_{2J} & 0 \end{pmatrix} \end{aligned}$$

**Theorem 33** *Given  $(\beta_g, \beta_b, \beta_p)$  satisfying Eq.47 with a strict inequality, define*

$$m(\beta_g, \beta_b, \beta_p) = \text{int} \left\{ \left( \frac{4\beta}{d+4\beta} - \frac{\beta_b}{d+2\beta_b} - \frac{\beta_p}{d+2\beta_p} \right) \left( 2 + \frac{d}{\beta_g} \right) + 1 \right\} + 1 \quad (49)$$

*to be the smallest integer such that  $\left( \frac{\log n}{n} \right)^{\frac{(m-1)\beta_g}{d+2\beta_g}} n^{-\frac{\beta_b}{d+2\beta_b} - \frac{\beta_p}{d+2\beta_p}} < n^{-\frac{4\beta}{d+4\beta}}$ , where  $\beta = \frac{\beta_b + \beta_p}{2}$ . Let  $\mathcal{K}_J, J, \widehat{\psi}_{3, \mathcal{K}_J}$  be as in Theorem 32 and define*

$$\begin{aligned} &\widehat{\psi}_{\mathcal{K}_J}^{eff}(\beta_g, \beta_b, \beta_p) \\ &= \widehat{\psi}_{3, \mathcal{K}_J} + \sum_{v=4}^{m(\beta_g, \beta_b, \beta_p)} (-1)^{v-1} \mathbb{H}_v^* + \sum_{s=1}^J \sum_{v=4}^{m(\beta_g, \beta_b, \beta_p)} (-1)^{v-1} \mathbb{G}(s, v) \\ &+ \sum_{v=4}^{m(\beta_g, \beta_b, \beta_p)} (-1)^{v-1} \mathbb{Q}_v \\ &= \mathbb{V}_{n,1} \left( H_1 \widehat{B} \widehat{P} + H_2 \widehat{B} + H_3 \widehat{P} + H_4 \right) - \mathbb{H}_2^* \\ &+ \sum_{v=3}^{m(\beta_g, \beta_b, \beta_p)} (-1)^{v-1} \mathbb{H}_v^* + \sum_{s=1}^J \sum_{v=3}^{m(\beta_g, \beta_b, \beta_p)} (-1)^{v-1} \mathbb{G}(s, v) + \sum_{v=3}^{m(\beta_g, \beta_b, \beta_p)} (-1)^{v-1} \mathbb{Q}_v \end{aligned}$$

Then

$$\begin{aligned}
& E \left( \widehat{\psi}_{\mathcal{K}_J}^{eff} (\beta_g, \beta_b, \beta_p) \right) - \psi (\theta) \\
&= O_p \left( \max \left[ \begin{aligned} & k_{-1}^{2\beta/d}, \left( \frac{\log n}{n} \right)^{-\frac{\beta_g}{d+2\beta_g}} k_{2s}^{-\beta_b/d} k_{2s+1}^{-\beta_p/d}, \left( \frac{\log n}{n} \right)^{-\frac{\beta_g}{d+2\beta_g}} k_{2s+1}^{-\beta_b/d} k_{2s}^{-\beta_p/d}, \\ & \left( \frac{\log n}{n} \right)^{-\frac{2\beta_g}{d+2\beta_g}} k_0^{-2\beta/d}, \left( \frac{\log n}{n} \right)^{-\frac{(m-1)\beta_g}{d+2\beta_g}} n^{-\frac{\beta_b}{d+2\beta_b} - \frac{\beta_p}{d+2\beta_p}}, \\ & \left( \frac{\log n}{n} \right)^{-\frac{2\beta_g}{d+2\beta_g}} \max_{1 \leq s \leq J} \left( k_{2s}^{-\beta_b/d} k_0^{-\beta_p/d}, k_0^{-\beta_b/d} k_{2s}^{-\beta_p/d} \right) \end{aligned} \right] \right) \\
&= O_p \left( \max \left[ \begin{aligned} & k_{-1}^{2\beta/d}, \left( \frac{\log n}{n} \right)^{-\frac{\beta_g}{d+2\beta_g}} k_{2s}^{-\beta_b/d} k_{2s+1}^{-\beta_p/d}, \left( \frac{\log n}{n} \right)^{-\frac{\beta_g}{d+2\beta_g}} k_{2s+1}^{-\beta_b/d} k_{2s}^{-\beta_p/d}, \\ & \left( \frac{\log n}{n} \right)^{-\frac{2\beta_g}{d+2\beta_g}} k_0^{-2\beta/d}, \left( \frac{\log n}{n} \right)^{-\frac{(m-1)\beta_g}{d+2\beta_g}} n^{-\frac{\beta_b}{d+2\beta_b} - \frac{\beta_p}{d+2\beta_p}} \end{aligned} \right] \right) \\
&= O_p \left( n^{-\frac{4\beta}{d+4\beta}} \right)
\end{aligned}$$

**Theorem 34** and

$$\begin{aligned}
& Var \left( \widehat{\psi}_{\mathcal{K}_J}^{eff} (\beta_g, \beta_b, \beta_p) \right) \\
&\asymp \frac{k_{-1}}{n^2} + \sum_{s=0}^J \frac{k_{2s} k_{2s-1}}{n^3} + \frac{k_{2J+1}^2}{n^3} \asymp n^{-\frac{8\beta}{d+4\beta}}
\end{aligned}$$

**Inference:** Elsewhere, we prove that  $\widehat{\psi}_{\mathcal{K}_J}^{eff} (\beta_g, \beta_b, \beta_p)$  is asymptotically normal.

Here, to avoid the problem of unknown 'constants' for confidence interval construction that we discussed in Section 3.2.5, we will construct nearly optimal rather than optimal confidence intervals. We suppose that Eq. (47) holds with strict equality for the  $(\beta_g, \beta_b, \beta_p)$  associated with the parameter space  $\Theta$ . Then there exists  $\epsilon > 0$  such that for all  $0 < \sigma < \epsilon$ ,  $(\beta_g, \beta_b - \sigma, \beta_p - \sigma)$  satisfies Eq (47) with strict equality,

$$\sup_{\theta \in \Theta} \left[ \frac{E_{\theta} \left[ \widehat{\psi}_{\mathcal{K}_J}^{eff} (\beta_g, \beta_b - \sigma, \beta_p - \sigma) | \widehat{\theta} \right]}{Var_{\theta} \left[ \widehat{\psi}_{\mathcal{K}_J}^{eff} (\beta_g, \beta_b - \sigma, \beta_p - \sigma) | \widehat{\theta} \right]} \right] = o_p(1)$$

and

$$\sup_{\theta \in \Theta} \left\{ \text{Var}_{\theta} \left[ \widehat{\psi}_{\mathcal{K}_J}^{eff} (\beta_g, \beta_b - \sigma, \beta_p - \sigma) | \widehat{\theta} \right] \right\} \asymp n^{-\frac{8(\beta-\sigma)}{d+4(\beta-\sigma)}}.$$

Let  $\widehat{\mathbb{W}} \left[ \widehat{\psi}_{\mathcal{K}_J}^{eff} (\beta_g, \beta_b, \beta_p) \right]$  be a uniformly consistent estimator of (the properly standardized)  $\text{Var}_{\theta} \left[ \widehat{\psi}_{\mathcal{K}_J}^{eff} (\beta_g, \beta_b, \beta_p) | \widehat{\theta} \right]$  constructed in the same manner as in Section 3.2.5. Then, for all  $\sigma < \epsilon$ ,

$$\left\{ \widehat{\psi}_{\mathcal{K}_J}^{eff} (\beta_g, \beta_b - \sigma, \beta_p - \sigma) - \psi(\theta) \right\} \left( \widehat{\mathbb{W}} \left[ \widehat{\psi}_{\mathcal{K}_J}^{eff} (\beta_g, \beta_b - \sigma, \beta_p - \sigma) \right] \right)^{-1}$$

converges uniformly in  $\theta \in \Theta$  to a  $N(0, 1)$ . Moreover,

$$\widehat{\psi}_{\mathcal{K}_J}^{eff} (\beta_g, \beta_b - \sigma, \beta_p - \sigma) \pm z_{\alpha} \widehat{\mathbb{W}} \left[ \widehat{\psi}_{\mathcal{K}_J}^{eff} (\beta_g, \beta_b - \sigma, \beta_p - \sigma) \right]$$

is a conservative uniform asymptotic  $(1 - \alpha)$  confidence interval for  $\psi(\theta)$  with diameter of the order of  $n^{-\frac{4(\beta-\sigma)}{d+4(\beta-\sigma)}}$ .

**Remark 35** *If Eq.47 holds with an equality and  $\mathcal{K}_J, J, \widehat{\psi}_{3, \mathcal{K}_J}$  are as in the final paragraph of the preceding subsection then the proof of Theorem 33 in the appendix implies  $\widehat{\psi}_{\mathcal{K}_J}^{eff} (\beta_g, \beta_b, \beta_p) - \psi(\theta) = O_p \left( (\log n) n^{-\frac{4\beta}{d+4\beta}} \right)$*

## 5 Adaptive Confidence Intervals for Regression and Treatment Effect Functions with unknown marginal of $X$ :

In this section we describe how to construct adaptive confidence intervals (i) for a regression function  $b(X) = E[Y|X]$  when the marginal of  $X$  is unknown and (ii) for

the treatment effect function and optimal treatment regime in a randomized clinical trial.

### 5.1 Regression Functions:

**Example 1a continued:** Consider the case  $b = p$ ,  $O = (Y, X)$  with  $b(X) = E(Y|X)$ . As usual, we assume for all  $\theta \in \Theta$ ,  $b(\cdot)$  and the density  $g(\cdot)$  of  $X$  are contained in known Hölder balls  $H(\beta_b, C_b)$  and  $H(\beta_g, C_g)$ . Redefine  $\psi(\theta) \equiv E_\theta \left[ \left( b(X) - \hat{b}(X) \right)^2 \right]$  where  $\hat{b}(\cdot)$  is an adaptive estimate of  $b(\cdot)$  from the training sample and expectations and probabilities remain conditional on the training sample. Adaptivity of  $\hat{b}(\cdot)$  implies that if  $b(\cdot) \in \theta$  is also contained in a smaller Hölder ball  $H(\beta^*, C)$ ,  $\beta^* > \beta_b, C < C_b$ , then  $\hat{b}(\cdot)$  will converge to  $b(\cdot)$  under  $F(\cdot, \theta)$  at rate  $O_p \left( n^{-\frac{\beta^*/d}{1+2\beta^*/d}} \right)$ . Robins and van der Vaart (2006) showed that, when the marginal density  $g(x)$  of  $X$  is known, the key to constructing optimal (rate) adaptive confidence balls for  $b(X)$  was to find a rate optimal estimator of  $E_\theta \left[ \left( b(X) - \hat{b}(X) \right)^2 \right]$ . We shall show that their approach fails when the marginal of  $X$  is unknown, but that a modification described below succeeds. Specifically, if  $b(\cdot) \in \theta$  lies in a smaller Hölder ball  $H(\beta^*, C)$ ,  $\beta^* > \beta_b, C < C_b$ , our modification results in honest asymptotic confidence balls under  $F(\cdot, \theta)$ ,  $\theta \in \Theta$ , whose diameter is (essentially) of the same order  $O_p \left( \max \left\{ n^{-\frac{\beta^*/d}{1+2\beta^*/d}}, n^{-\frac{2\beta_b}{d+4\beta_b}} \right\} \right)$  as the diameter of Robins and van der Vaart's optimal adaptive region or ball, provided either (i)  $\beta_b/d > 1/4$  and  $\beta_g/d > 0$  or (ii)  $\beta_b/d \leq 1/4$  and eq.(37) holds with  $\beta = \beta_b$ .

This order is the maximum of the minimax rate  $n^{-\frac{\beta^*/d}{1+2\beta^*/d}}$  of convergence of  $\widehat{b}(X)$  to  $b(X)$  were  $b(X)$  known to lie in  $H(\beta^*, C)$  and the square root of the minimax rate of convergence of an estimator of  $E_\theta \left[ \left( b(X) - \widehat{b}(X) \right)^2 \right]$  in the larger model  $\mathcal{M}(\Theta)$  with  $b(\cdot)$  and  $g(\cdot)$  only known to lie in  $H(\beta_b, C_b)$  and  $H(\beta_g, C_g)$ .

The case where  $\beta_b/d \leq 1/4$  and eq.(37) does not hold will be considered elsewhere.

Now, since  $E_\theta \left[ \widehat{b}(X) b(X) \right] = E_\theta \left[ \widehat{b}(X) Y \right]$ ,

$$\psi(\theta) \equiv E_\theta \left[ \left( b(X) - \widehat{b}(X) \right)^2 \right] = E_\theta \left[ \{b(X)\}^2 \right] - 2E_\theta \left[ \widehat{b}(X) b(X) \right] + E_\theta \left[ \left\{ \widehat{b}(X) \right\}^2 \right]$$

has first order influence function  $\mathbb{IF}_{1,\psi}(\theta) = \mathbb{V}[H(b, b) - \psi(\theta)]$  where

$$H(b, b) = b^2(X) + 2b(X)[Y - b(X)] - 2\widehat{b}(X)Y + \widehat{b}^2(X),$$

so  $H_1 = -1, H_2 = H_3 = Y, H_4 = -2\widehat{b}(X)Y + \widehat{b}^2(X)$ . Thus  $H(b, b)$  for  $E_\theta[b(X)^2]$  differs from  $H(b, b)$  for  $E_\theta \left[ \left( b(X) - \widehat{b}(X) \right)^2 \right]$  only in  $H_4$ . Since the truncation bias  $\widetilde{\psi}_k(\theta) - \psi(\theta)$ , higher order influence functions of  $\widetilde{\psi}_k(\theta)$  and estimation bias do not depend on  $H_4$ , it follows that  $TB_k(\theta), \mathbb{IF}_{jj, \widetilde{\psi}_k}(\theta), \widehat{\mathbb{W}}_{jj, \widetilde{\psi}_k}^2$ , and  $EB_m(\theta)$  are identical for  $\psi(\theta) \equiv E_\theta \left[ \left( b(X) - \widehat{b}(X) \right)^2 \right]$  and  $\psi(\theta) \equiv E_\theta[b(X)^2]$ . In contrast,  $IF_{1,\psi}(\widehat{\theta})$  is identically zero for  $\psi(\theta) \equiv E_\theta \left[ \left( b(X) - \widehat{b}(X) \right)^2 \right]$  but not for  $\psi(\theta) \equiv E_\theta[b(X)^2]$ . Thus, by Theorem (27), for  $\psi(\theta) \equiv E_\theta \left[ \left( b(X) - \widehat{b}(X) \right)^2 \right]$ ,  $\text{var}_\theta \left[ \widehat{\psi}_{m, \widetilde{\psi}_k} \right] \asymp \frac{1}{n} \left( \frac{k}{n} \right)^{m-1}$  if  $k > n$  and  $m > 1$ , and  $\text{var}_\theta \left[ \widehat{\psi}_{m, \widetilde{\psi}_k} \right] = 0$  if  $k \leq n$  and  $m = 1$ . In the case when  $k \leq n$  and  $m > 1$ , by the Hoeffding decomposition,

$$\text{var}_\theta \left[ \widehat{\psi}_{m, \widetilde{\psi}_k} \right] = \text{var}_\theta \left( \sum_{s=1}^m \left( \mathbb{D}_s^{(\widehat{\psi}_{m, \widetilde{\psi}_k})}(\theta) \right) \right)$$

where  $\mathbb{D}_s^{(\hat{\psi}_m, \tilde{\psi}_k)}$  is a sth order degenerate U-statistic. Further by Theorem (27), we have

$$\text{var}_\theta [\hat{\psi}_{m, \tilde{\psi}_k}] \asymp \max \left( \text{var}_\theta \left( \mathbb{D}_1^{(\hat{\psi}_m, \tilde{\psi}_k)} \right), \text{var}_\theta \left( \mathbb{D}_2^{(\hat{\psi}_m, \tilde{\psi}_k)} \right) \right)$$

as  $\text{var}_\theta \left( \mathbb{D}_s^{(\hat{\psi}_m, \tilde{\psi}_k)} \right) \asymp \frac{1}{n} \left( \frac{k}{n} \right)^{s-1} = o \left( \frac{k}{n^2} \right)$  for any  $s > 2$ . Moreover,

$$\text{var}_\theta \left( \mathbb{D}_1^{(\hat{\psi}_m, \tilde{\psi}_k)} \right) \asymp \frac{\left\| b(X) - \hat{b}(X) \right\|_2^2}{n}$$

since the kernel of  $\mathbb{D}_1^{(\hat{\psi}_m, \tilde{\psi}_k)}$  is of order  $O_p \left( \left\| b(X) - \hat{b}(X) \right\|_2 \right)$ . In summary

$$\begin{aligned} \text{var}_\theta [\hat{\psi}_{m, \tilde{\psi}_k}] &\asymp \max \left( \frac{\left\| b(X) - \hat{b}(X) \right\|_2^2}{n}, \frac{k}{n^2} \right) \\ &= \max \left( n^{-\frac{2\beta_b/d}{1+4\beta_b/d}-1}, \frac{k}{n^2} \right) \end{aligned}$$

if  $k \leq n$  and  $m > 1$  (In contrast, for  $\psi(\theta) \equiv E_\theta [b(X)^2]$ ,  $\text{var}_\theta [\hat{\psi}_{m, \tilde{\psi}_k}] \asymp \max \left( \frac{1}{n}, \frac{k}{n^2} \right) =$

$\frac{1}{n}$  if  $k \leq n$ ). Thus if  $\beta_b/d > 1/4$ , (i)  $\hat{\psi}_{m_{opt}, k_{opt}(m_{opt})}$  has  $k_{opt}(m_{opt})$  of  $O \left( n^{\frac{2}{1+4\beta_b/d}} \right)$ ,

where  $n^{\frac{2}{1+4\beta_b/d}} < n$  comes from equating the order  $k^{-4\beta_b/d}$  of  $TB_k^2(\theta)$  to the order  $k/n^2 = n^{-\frac{8\beta_b/d}{1+4\beta_b/d}} \ll n^{-1}$  of the variance ( $n^{-\frac{2\beta_b/d}{1+4\beta_b/d}-1} \ll n^{-\frac{8\beta_b/d}{1+4\beta_b/d}}$  for  $\forall \beta_b > 0$ )

and (ii)  $m_{opt}$  is the smallest integer  $m$  such that the order  $n^{-\left(\frac{(m-1)\beta_g}{2\beta_g+d} + \frac{2\beta_b}{d+2\beta_b}\right)}$  of

$EB_m = O_p \left( n^{-\left(\frac{(m-1)\beta_g}{2\beta_g+d} + \frac{2\beta_b}{d+2\beta_b}\right)} \right)$  is less than the order  $n^{-\frac{4\beta_b/d}{1+4\beta_b/d}}$  of the standard error.

It follows that, for  $\beta_b/d > 1/4$ , in contrast to  $\psi(\theta) \equiv E_\theta [b(X)^2]$ , we can estimate

$\psi(\theta) \equiv E_\theta \left[ \left( b(X) - \hat{b}(X) \right)^2 \right]$  at (the minimax) rate  $n^{-\frac{4\beta_b/d}{1+4\beta_b/d}}$  which is faster (i.e.,

less) than the usual parametric rate of  $n^{-1/2}$ .



When  $\beta_b/d < 1/4$ , the minimax rates for  $\psi(\theta) \equiv E_\theta \left[ \left( b(X) - \widehat{b}(X) \right)^2 \right]$  and  $\psi(\theta) \equiv E_\theta \left[ b(X)^2 \right]$  are identical and, when eq.(37) holds, it follows from Theorem 33 that  $\widehat{\psi}_{\mathcal{K}_J}^{eff}(\beta_g, \beta_b, \beta_b)$  achieves the minimax rate of  $n^{-\frac{4\beta_b/d}{1+4\beta_b/d}} \geq n^{-1/2}$ .

Henceforth assume either (i)  $\beta_b/d > 1/4$  or (ii)  $\beta_b/d < 1/4$  and eq.(37) holds. Pick an  $\epsilon$  so that eq.(37) holds for  $(\beta_g, \beta_b - \epsilon, \beta_p - \epsilon)$ . Let  $0 < \sigma < \epsilon$  and define  $\widehat{\psi}^* \equiv \widehat{\psi}(\sigma) = \widehat{\psi}_{m_{opt}, \{k_{opt}(m_{opt})\}^{1+\sigma}}$  and  $\widehat{\mathbb{W}}^* \equiv \widehat{\mathbb{W}}^*(\sigma) = \widehat{\mathbb{W}}_{m_{opt}, \tilde{\psi}_{\{k_{opt}(m_{opt})\}^{1+\sigma}}}$  if  $\beta_b/d > 1/4$  and  $\widehat{\psi}^* = \widehat{\psi}_{\mathcal{K}_J}^{eff}(\beta_g, \beta_b - \sigma, \beta_p - \sigma)$  and  $\widehat{\mathbb{W}}^* = \widehat{\mathbb{W}} \left[ \widehat{\psi}_{\mathcal{K}_J}^{eff}(\beta_g, \beta_b - \sigma, \beta_p - \sigma) \right]$  if  $\beta_b/d < 1/4$ . Note  $\widehat{\mathbb{W}}^*$  is  $O_p \left( n^{-\frac{4(\beta_b - \sigma)}{d+4(\beta_b - \sigma)}} \right)$  uniformly over  $\Theta$ , where  $\Theta$  is the parameter space with smoothness parameters  $(\beta_g, \beta_b)$ . Then, by eq.(37) and results in Section 4.1.2, as  $n \rightarrow \infty$ ,  $\inf_{\theta \in \Theta} \text{pr}_\theta \left[ \left\{ \widehat{\psi}^* - \psi(\theta) \right\} \geq -z_\alpha \widehat{\mathbb{W}}^* \right] \geq 1 - \alpha$ . Thus, if  $\psi(\theta)$  were a function of  $\theta$  only through  $b(\cdot)$  so  $\psi(\theta) = \psi(b)$ , the set

$$\left\{ b^*(\cdot); \psi(\theta) \leq \widehat{\psi}^* + z_\alpha \widehat{\mathbb{W}}^* \right\} \quad (50)$$

would be an uniform asymptotic  $(1 - \alpha)$  confidence region for  $b(\cdot)$ . However, for  $\psi(\theta) = E_\theta \left[ \left( b(X) - \widehat{b}(X) \right)^2 \right]$ , this approach fails because  $\psi(\theta)$  also depends on  $\theta$  through the unknown density  $g(x)$  of  $X$ . This approach succeeded in Robins and van der Vaart (2006) because  $g(x)$  was assumed known.

We consider two solutions. The first gives (near) optimal adaptive honest intervals. The second would give honest, but non-optimal, intervals. The first solution is to

replace  $\psi(\theta)$  with its empirical mean  $\psi_{emp}(b) \equiv \mathbb{V} \left[ \left\{ b(X) - \widehat{b}(X) \right\}^2 \right]$  in eq.(50).

$$\psi_{emp}(b) - \psi(\theta) = O_p \left( \left[ \left\{ b(X) - \widehat{b}(X) \right\}^2 \right] n^{-1/2} \right) = O_p \left( n^{-\left(\frac{2\beta_b}{d+2\beta_b} + \frac{1}{2}\right)} \right)$$

uniformly in  $\theta \in \Theta$ . It is straightforward to check that for all  $\beta_b > 0$ ,  $n^{-\left(\frac{2\beta_b}{d+2\beta_b} + \frac{1}{2}\right)} \ll n^{-\frac{4\beta_b/d}{1+4\beta_b/d}}$ . Thus, for  $\sigma < \epsilon$ ,  $\left\{ \widehat{\psi}^* - \psi_{emp}(b) \right\} / \left\{ \widehat{\psi}^* - \psi(\theta) \right\} = 1 + o_p(1)$  uniformly over  $\theta \in \Theta$ , so  $\inf_{\theta \in \Theta} \Pr_{\theta} \left[ \left\{ \widehat{\psi}^* - \psi_{emp}(b) \right\} \geq -z_{\alpha} \widehat{\mathbb{W}}^* \right] \geq 1 - \alpha$  and

$$\left\{ b^*(\cdot); \mathbb{V} \left[ \left\{ b^*(X) - \widehat{b}(X) \right\}^2 \right] \leq \widehat{\psi}^* + z_{\alpha} \widehat{\mathbb{W}}^* \right\} \quad (51)$$

is a uniform asymptotic  $(1 - \alpha)$  confidence region for  $b(\cdot)$ . Moreover, if  $b(\cdot) \in \theta$  lies in a smaller Hölder ball  $H(\beta^*, C)$ ,  $\beta^* > \beta_b$ ,  $C < C_b$ , then, under  $F(\cdot, \theta)$ , the diameter

$$\begin{aligned} \left\{ \widehat{\psi}^* + z_{\alpha} \widehat{\mathbb{W}}^* \right\}^{1/2} &= \left\{ \psi(\theta) + O_p \left( n^{-\frac{4(\beta_b - \sigma)}{d+4(\beta_b - \sigma)}} \right) \right\}^{1/2} \\ &= O_p \left( \max \left\{ n^{-\frac{2\beta^*/d}{1+2\beta^*/d}}, n^{-\frac{4(\beta_b - \sigma)}{d+4(\beta_b - \sigma)}} \right\} \right)^{1/2} \\ &= O_p \left( \max \left\{ n^{-\frac{\beta^*/d}{1+2\beta^*/d}}, n^{-\frac{2(\beta_b - \sigma)}{d+4(\beta_b - \sigma)}} \right\} \right) \end{aligned}$$

since  $\psi(\theta) = O_p \left( n^{-\frac{2\beta^*/d}{1+2\beta^*/d}} \right)$  and  $\widehat{\psi}^* - \psi(\theta)$  and  $\widehat{\mathbb{W}}^*$  are  $O_p \left( n^{-\frac{4(\beta_b - \sigma)}{d+4(\beta_b - \sigma)}} \right)$ .

The second, non-optimal, solution would be to replace the functional  $\psi(\theta) \equiv E_{\theta} \left[ \left( b(X) - \widehat{b}(X) \right)^2 \right]$  with  $\psi(b) = \int \left\{ b(x) - \widehat{b}(x) \right\}^2 dx$ . The functional  $\psi(b)$  is the first functional we have considered that is not in our doubly robust class of functionals. Arguing as above, if we can construct an asymptotically normal higher order  $U$ -statistic estimator  $\widehat{\psi}^*$  that converges to  $\psi(b)$  at rate  $n^{-\omega}$  on  $\mathcal{M}(\Theta)$  and a consistent

estimator  $\widehat{\mathbb{W}}^*$  of its standard error, then  $\left\{b^*(\cdot); \int \left\{b(x) - \widehat{b}(x)\right\}^2 dx \leq \widehat{\psi}^* + z_\alpha \widehat{\mathbb{W}}^*\right\}$  would be an honest adaptive confidence interval of diameter  $O_p\left(\max\left\{n^{-\frac{\beta^*/d}{1+2\beta^*/d}}, n^{-\omega/2}\right\}\right)$ .

We conjecture, based on arguments given elsewhere, that the minimax rate for estimation of  $\psi(b) = \int \left\{b(x) - \widehat{b}(x)\right\}^2 dx$  exceeds  $O_p\left[n^{-\frac{4\beta_b}{d+4\beta_b}}\right]$  whenever  $\frac{\beta_g/d}{2\beta_g/d+1} < \frac{\beta/d}{(1+4\beta/d)(1+2\beta/d)}$ . Since  $\frac{\beta/d}{(1+4\beta/d)(1+2\beta/d)} > \frac{1-4\beta/d}{1+4\beta/d}\beta/d$  for all  $\beta > 0$ , it follows that, when the marginal of  $X$  is unknown and  $\frac{\beta/d}{(1+4\beta/d)(1+2\beta/d)} > \frac{\beta_g/d}{2\beta_g/d+1} > \frac{1-4\beta/d}{1+4\beta/d}\beta/d$ , intervals based on  $\mathbb{V}\left[\left\{b^*(X) - \widehat{b}(X)\right\}^2\right]$  will, but intervals based on  $\int \left\{b(x) - \widehat{b}(x)\right\}^2 dx$  will not, have diameter of the same order as the optimal interval with the marginal of  $X$  known.

## 5.2 Treatment Effect Functions in a Randomized Trial

**Example 4 continued:** Consider the case  $b = p$ ,  $Y = Y^*$  wpl so we have data  $O = \{Y, A, X\}$ , where  $A$  is a binary treatment,  $Y$  is the response, and  $X$  is a vector of pre-randomization covariates. The randomization probabilities  $\pi_0(X) = P(A = 1|X)$  are known by design and  $b(x) = E_\theta(Y|A = 1, X = x) - E_\theta(Y|A = 0, X = x)$  is the average treatment effects function. For  $\theta \in \Theta$ ,  $b(\cdot)$  and the density  $g(\cdot)$  of  $X$  are contained in known Hölder balls  $H(\beta_b, C_b)$  and  $H(\beta_g, C_g)$ . Suppose we have an adaptive estimator  $\widehat{b}(\cdot)$  of  $b(\cdot)$  based on the training sample constructed as described below. Now, since  $E_\theta[\widehat{b}(X)b(X)] = E_\theta[\widehat{b}(X)Y|A = 1] - E_\theta[\widehat{b}(X)Y|A = 0]$  has influence function  $\frac{A}{\pi_0(X)}Y\widehat{b}(X) - \frac{1-A}{1-\pi_0(X)}Y\widehat{b}(X) - E_\theta[\widehat{b}(X)b(X)] = (A - \pi_0(X))\sigma_0^{-2}(X)Y\widehat{b}(X) -$

$E_\theta [\widehat{b}(X) b(X)]$ , where  $\sigma_0^2(X) = \pi_0(X) \{1 - \pi_0(X)\}$ ,  $\psi(\theta) \equiv E_\theta \left[ \left( b(X) - \widehat{b}(X) \right)^2 \right]$  has first order influence functions, indexed by arbitrary functions  $c(x)$ ,  $\mathbb{IF}_{1,\psi}(\theta, c) \equiv \mathbb{IF}_{1,\psi}(\theta) = \mathbb{V} [H(b, b) - \psi(\theta)]$  with

$$H_1 = 1 - 2A \{A - \pi_0(X)\} \sigma_0^{-2}(X),$$

$$H_2 = H_3 = \{A - \pi_0(X)\} \sigma_0^{-2}(X) Y,$$

$$H_4 = \{A - \pi_0(X)\} c(X) - 2(A - \pi_0(X)) \sigma_0^{-2}(X) Y \widehat{b}(X) + \widehat{b}^2(X)$$

Thus  $H(b, b)$  for  $E_\theta \left[ \left( b(X) - \widehat{b}(X) \right)^2 \right]$  differs from  $H(b, b)$  for  $\psi(\theta) \equiv E_\theta [b(X)^2]$  only in  $H_4$ . It follows that all the properties of the confidence ball 51 for  $b(\cdot) = E_\theta(Y|X = \cdot)$  in the setting of the last subsection remain true for  $b(\cdot) = E_\theta(Y|A = 1, X = \cdot) - E_\theta(Y|A = 0, X = \cdot)$  in the setting of this subsection.

Now define  $d_{b^*}(x) = I[b^*(x) > 0]$ . Then it then follows that an honest  $1 - \alpha$  uniform asymptotic confidence set for the optimal treatment regime  $d_{opt}(\cdot) = I[b(\cdot) > 0]$  is given by  $\left\{ d_{b^*}(\cdot); \mathbb{V} \left[ \left\{ b^*(X) - \widehat{b}(X) \right\}^2 \right] \leq \widehat{\psi}^* + z_\alpha \widehat{\mathbb{W}}^* \right\}$ .

**Adaptive Estimator of The Treatment Effect Function:** One among many approaches to constructing a rate-adaptive estimator of  $b(\cdot)$  is as follows. Split the training sample into two random subsamples - a candidate estimator subsample of size  $n_c$  and a validation subsample of size  $n_v$ , where both  $n_c/n$  and  $n_v/n$  are bounded away from 0 as  $n \rightarrow \infty$ . Noting that  $0 = \mathbb{E}_\theta [\{Y - Ab(X)\} q(X) \{A - \pi_0(X)\}]$  for all  $q(\cdot)$ , we construct candidate estimators of  $b(\cdot)$  as follows. For  $s = 1, 2, \dots, n - 1$ ,

let  $\widehat{\boldsymbol{\varkappa}}_s$  be the solution, if any, to the  $s$  equations

$$0 = \mathbb{P}_c \left[ \left\{ Y - A \boldsymbol{\varkappa}_s^T \overline{\boldsymbol{\varphi}}_s(X) \right\} \overline{\boldsymbol{\varphi}}_s(X) \{A - \pi_0(X)\} \right]$$

where  $\varphi_1(X), \varphi_2(X), \dots$  is a complete basis wrt to Lebesgue measure in  $R^d$  that provides optimal rate approximation for Hölder balls and  $\mathbb{P}_c$  is the empirical measure for the candidate estimator subsample. Our candidates for  $b(X)$  are the  $\widehat{b}^{(s)}(X) = \overline{\boldsymbol{\varphi}}_s(X)^T \widehat{\boldsymbol{\varkappa}}_s$ . Robins (2004) proved that  $b(\cdot)$  is the unique function  $b^*(\cdot)$  minimizing  $Risk(b^*) \equiv E_\theta [\sigma_0^{-2}(X) \{Y - [A - \pi_0(X)] b^*(X)\}^2]$ . In fact, the candidate  $\widehat{b}^{(s)}(X)$  in our set for which  $Risk(\widehat{b}^{(s)})$  is smallest is also the candidate that minimizes  $E \left[ \left( b(X) - \widehat{b}^{(s)}(X) \right)^2 \right]$  since  $Risk(\widehat{b}^{(s)}) - Risk(b) = E \left[ \left( b(X) - \widehat{b}^{(s)}(X) \right)^2 \right]$ . Specifically,

$$\begin{aligned} & E \left[ \begin{array}{l} \sigma_0^{-2}(X) \left\{ Y - [A - \pi_0(X)] \widehat{b}^{(s)}(X) \right\}^2 \\ - \sigma_0^{-2}(X) \left\{ Y - [A - \pi_0(X)] b(X) \right\}^2 \end{array} \right] \\ &= E \left[ \begin{array}{l} \sigma_0^{-2}(X) (A - \pi_0(X)) \left( b(X) - \widehat{b}^{(s)}(X) \right) \times \\ \left( 2(Ab(X) - E(Y|A=0, X)) - (A - \pi_0(X)) (b(X) + \widehat{b}^{(s)}(X)) \right) \end{array} \right] \\ &= E \left( \sigma_0^{-2}(X) (A - \pi_0(X)) A \left( b(X) - \widehat{b}^{(s)}(X) \right)^2 \right) \\ &= E \left[ \left( b(X) - \widehat{b}^{(s)}(X) \right)^2 \right] \end{aligned}$$

We use these results to select among our candidates by cross-validation. Let  $\widehat{b}(\cdot)$  be the  $\widehat{b}^{(s)}(\cdot)$  minimizing  $\mathbb{P}_v \left[ \sigma_0^{-2}(X) \left\{ Y - [A - \pi_0(X)] \widehat{b}^{(s)}(X) \right\}^2 \right]$  over  $s = 1, 2, \dots, n-1$ , where  $\mathbb{P}_v$  is the validation subsample empirical measure. If  $b(\cdot)$  were known to lie in

a Hölder ball  $H(\beta, C)$ , it is easy to check that the candidate  $\widehat{b}^{(s)}(\cdot)$  with  $s = \lfloor n^{\frac{1}{2\beta+1}} \rfloor$  obtains the optimal rate of  $n^{\frac{-\beta}{2\beta+1}}$  for estimating  $E \left[ \left( b(X) - \widehat{b}^{(s)}(X) \right)^2 \right]$ . Since the number of candidates at sample size  $n$  is less than  $n$ , it then follows at once from van der Laan and Dudoit's (2003) results on model selection by cross validation that  $\widehat{b}(\cdot)$  is adaptive over Hölder balls.

## 6 Testing, Confidence Sets, and Implicitly Defined Functionals:

In Example 1c of section 3.1, we considered the following problem. We were given a functional  $\psi(\tau, \theta)$  indexed by a real number  $\tau$  and the parameter  $\theta \in \Theta$ . The implicitly defined-functional  $\tau(\theta)$  was the assumed unique solution to  $0 = \psi(\tau, \theta)$ . We noted that a  $(1 - \alpha)$  confidence set for  $\tau(\theta)$  is the set of  $\tau$  such that a  $(1 - \alpha)$  CI interval for  $\psi(\tau, \theta)$  contains 0. In the following subsection we derive the width of the confidence set for  $\tau(\theta)$ . We then generalize the problem in the second subsection by introducing the notions of the testing tangent space, a testing influence function, and the higher order efficient testing score. In the final subsection, we show how the two earlier subsections are related.

### 6.1 Confidence Intervals for Implicitly Defined Functionals:

To derive the order of the length of the confidence interval for the parameter  $\tau(\theta)$  in Example 1c, we can use the next theorem as follows. Assume eq (37) holds and

$\beta \leq 1/4$ . Then we can take the estimator  $\widehat{\psi}(\tau)$  and rate  $n^{-\gamma}$  in the theorem to be the estimator  $\widehat{\psi}_{\mathcal{K}_J}^{eff}$  and rate  $n^{-\frac{4\beta}{4\beta+1}+\sigma}$  for a very small positive  $\sigma$  and conclude that the length of the confidence interval for  $\tau(\theta)$  in Example 1c to be  $O_p\left(n^{-\frac{4\beta}{4\beta+1}+\sigma}\right)$ .

**Theorem 36 :** *Suppose for an estimator  $\widehat{\psi}(\tau)$  and functional  $\psi(\tau, \theta)$ , there is a scale estimator  $\widehat{\mathbb{W}}(\tau)$  such that  $n^\gamma \widehat{\mathbb{W}}(\tau) \rightarrow w(\tau, \theta)$  in  $\theta$ -probability,  $w(\tau, \theta) > c^* > 0$  and  $\left(\widehat{\psi}(\tau) - \psi(\tau, \theta)\right) / \widehat{\mathbb{W}}(\tau)$  converges in law to  $N(0, 1)$  uniformly for  $\theta \in \Theta$ ,  $\tau \in \{\tau(\theta); \theta \in \Theta\}$ . Then, (i) with  $z_\alpha$  the  $\alpha$ -quantile and  $\Phi(\cdot)$  the CDF of a  $N(0, 1)$ , the confidence set  $\mathcal{C}_n = \left\{\tau; -z_{1-\alpha/2} < \frac{\widehat{\psi}(\tau)}{\widehat{\mathbb{W}}(\tau)} < z_{1-\alpha/2}\right\}$  is a uniform asymptotic  $1 - \alpha$  confidence set for the (assumed) unique solution  $\tau(\theta)$  to  $\psi(\tau, \theta) = 0$ ; (ii) the probability under  $\theta$  that a sequence  $\tau = \tau_n$  satisfying  $\psi(\tau_n, \theta) = a_n n^{-\rho}$ ,  $a_n \rightarrow a \neq 0$ , is contained in  $\mathcal{C}_n$  converges to 1 when  $\rho > \gamma$ , is  $o(1)$  when  $\rho < \gamma$ , and converges to  $\Phi\left(z_{1-\alpha/2} - \frac{a}{w(\tau(\theta), \theta)}\right) - \Phi\left(-z_{1-\alpha/2} - \frac{a}{w(\tau(\theta), \theta)}\right)$  when  $\rho = \gamma$ . (iii) If  $\psi(\tau, \theta)$  is uniformly twice continuously differentiable in  $\tau$  and  $0 < \sigma < |\psi_\tau(\tau(\theta), \theta)| < c$  and  $|\psi_{\tau^2}(\tau(\theta), \theta)| < c$  for constants  $(\sigma, c)$ , then (ii) holds for a sequence  $\tau = \tau_n$  satisfying  $\tau_n - \tau(\theta) = \{\psi_\tau(\tau(\theta), \theta)\}^{-1} a_n n^{-\rho}$ ,  $a_n \rightarrow a \neq 0$ ,  $\rho > 0$ .*

**Proof.** (i): That  $\mathcal{C}_n$  is a uniform asymptotic  $1 - \alpha$  confidence set is immediate. (ii):

Now

$$\begin{aligned}
& Pr_\theta \left\{ z_{1-\alpha/2} > \frac{\widehat{\psi}(\tau_n)}{\widehat{\mathbb{W}}(\tau_n)} > -z_{1-\alpha/2} \right\} \\
&= Pr_\theta \left\{ z_{1-\alpha/2} - \frac{\psi(\tau_n, \theta)}{\widehat{\mathbb{W}}(\tau_n)} > \frac{\widehat{\psi}(\tau_n) - \psi(\tau_n, \theta)}{\widehat{\mathbb{W}}(\tau_n)} > -z_{1-\alpha/2} - \frac{\psi(\tau_n, \theta)}{\widehat{\mathbb{W}}(\tau_n)} \right\} \\
&\xrightarrow{n \rightarrow \infty} \Phi \left( z_{1-\alpha/2} - \lim_{n \rightarrow \infty} \frac{n^\gamma \psi(\tau_n, \theta)}{n^\gamma \widehat{\mathbb{W}}(\tau_n)} \right) - \Phi \left( -z_{1-\alpha/2} - \lim_{n \rightarrow \infty} \frac{n^\gamma \psi(\tau_n, \theta)}{n^\gamma \widehat{\mathbb{W}}(\tau_n)} \right) \\
&= \Phi \left( z_{1-\alpha/2} - \frac{a \lim_{n \rightarrow \infty} n^{\gamma-\rho}}{w(\tau(\theta), \theta)} \right) - \Phi \left( -z_{1-\alpha/2} - \frac{a \lim_{n \rightarrow \infty} n^{\gamma-\rho}}{w(\tau(\theta), \theta)} \right).
\end{aligned}$$

(iii): Since  $\psi(\tau_n, \theta) = \psi_\tau(\tau(\theta), \theta)(\tau_n - \tau(\theta)) + \frac{1}{2}\psi_{\tau^2}(\tau^*(\theta), \theta)(\tau_n - \tau(\theta))^2$  for some  $\tau^*(\theta)$  between  $\tau(\theta)$  and  $\tau$ , we have that  $\psi(\tau_n, \theta) = a_n n^{-\rho} + o_p(a_n n^{-\rho}) = a_n(1 + o_p(1))n^{-\rho}$  satisfies the assumption in (ii). ■

**Remark:** Under some further regularity conditions, the solution  $\tilde{\tau}$  to  $0 = \tilde{\psi}(\tau)$  is asymptotically normal with mean  $\tau(\theta)$  and variance  $\psi_\tau^{-2}(\tau, \theta) [\{w(\tau(\theta), \theta)\}^2]$  uniformly over  $\theta \in \Theta, \tau \in \{\tau(\theta); \theta \in \Theta\}$ .

## 6.2 Testing influence functions and a higher order efficient score

In the following, we repeatedly use definitions from Sec. 2, which might usefully be reviewed at this point.

**Definition 37**  *$m^{\text{th}}$  order testing nuisance tangent space, testing tangent space, testing influence functions, efficient score, efficient infor-*



**mation, and efficient testing variance:** Given a model  $\mathcal{M}(\Theta)$  with parameter space  $\Theta$  and a functional  $\tau(\theta)$ , define  $\mathcal{M}(\Theta(\tau^\dagger))$  to be the submodel with parameter space  $\Theta(\tau^\dagger) \equiv \Theta \cap \{\theta; \tau(\theta) = \tau^\dagger\}$ . Thus  $\mathcal{M}(\Theta(\tau^\dagger))$  is the submodel with  $\tau(\theta)$  equal to  $\tau^\dagger$ . Define, for  $\theta \in \Theta(\tau^\dagger)$ , the  $m$ th order (i) testing nuisance tangent space  $\Gamma_m^{nuis, test}(\theta, \tau^\dagger)$  to be the  $m^{th}$  order tangent space for the submodel  $\mathcal{M}(\Theta(\tau^\dagger))$ , (ii) testing tangent space  $\Gamma_m^{test}(\theta, \tau^\dagger)$  to be the closed linear span of  $\mathbb{IF}_{1, \tau(\cdot)}(\theta) \cup \Gamma_m^{nuis, test}(\theta, \tau^\dagger)$ , (iia) set  $\Gamma_m^{nuis, test, \perp}(\theta, \tau^\dagger) \equiv \{\mathbb{IF}_{m, \tau(\cdot)}^{test}\}$  of testing influence functions to be the orthocomplement of  $\Gamma_m^{nuis, test}(\theta, \tau^\dagger)$  in  $\mathcal{U}_m(\theta)$ , (iib) set  $\Gamma_m^{std, nuis, test, \perp}(\theta, \tau^\dagger) \equiv \{\mathbb{IF}_{m, \tau(\cdot)}^{std, test}\}$  of standardized testing influence functions to be

$$\left\{ \mathbb{IF}_{m, \tau(\cdot)}^{std, test} \in \Gamma_m^{nuis, test, \perp}(\theta, \tau^\dagger); E_\theta \left[ \mathbb{IF}_{m, \tau(\cdot)}^{std, test} \mathbb{IF}_{1, \tau(\cdot)}^{eff}(\theta) \right] = \text{var}_\theta \left[ \mathbb{IF}_{1, \tau(\cdot)}^{eff}(\theta) \right] \right\},$$

(iv) efficient testing score  $\mathbb{ES}_m^{test}(\theta) \equiv \mathbb{ES}_{m, \tau(\cdot)}^{test}(\theta) \in \Gamma_m^{test}(\theta, \tau^\dagger)$  to be

$$\mathbb{ES}_{m, \tau(\cdot)}^{test}(\theta) = \mathbb{ES}_1^{test} - \Pi_\theta \left[ \mathbb{ES}_1^{test} | \Gamma_m^{nuis, test}(\theta, \tau^\dagger) \right] \equiv \Pi_\theta \left[ \mathbb{ES}_1^{test}(\theta) | \Gamma_m^{nuis, test, \perp}(\theta, \tau^\dagger) \right]$$

where  $\mathbb{ES}_1^{test}(\theta) \equiv \mathbb{ES}_{1, \tau(\cdot)}^{test}(\theta) \equiv \text{var}_\theta \left\{ \mathbb{IF}_{1, \tau(\cdot)}^{eff}(\theta) \right\}^{-1} \mathbb{IF}_{1, \tau(\cdot)}^{eff}(\theta)$ , (v) efficient testing information to be  $\text{var}_\theta \left\{ \mathbb{ES}_m^{test}(\theta) \right\}$ , and (vi) the efficient testing variance to be  $\left[ \text{var}_\theta \left\{ \mathbb{ES}_m^{test}(\theta) \right\} \right]^{-1}$ .

Further define, for  $\theta \in \Theta$ , the  $m$ th order (i) estimation nuisance tangent space  $\Gamma_m^{nuis}(\theta)$  to be  $\Gamma_m^{nuis}(\theta) \equiv \left\{ \mathbb{A}_m \in \Gamma_m(\theta); E \left[ \mathbb{A}_m \mathbb{IF}_{m, \tau(\cdot)}^{eff}(\theta) \right] = 0 \right\}$ , and (ii) efficient estimation variance to be  $\text{var}_\theta \left[ \mathbb{IF}_{m, \tau(\cdot)}^{eff}(\theta) \right]$ .

**Remark:** For  $m = 1$ , the testing and estimation nuisance tangent spaces  $\Gamma_m^{nuis, test}(\theta, \tau^\dagger)$  and  $\Gamma_m^{nuis}(\theta)$  are identical. However for  $m > 1$ ,  $\Gamma_m^{nuis, test}(\theta, \tau^\dagger)$  is generally a strict subset of  $\Gamma_m^{nuis}(\theta)$ . For example, if the model can be parametrized as  $\theta = (\tau, \rho)$  and  $\Theta$  is the product of the parameter spaces for  $\tau$  and  $\rho$ , the  $\Gamma_m^{nuis, test}(\theta, \tau^\dagger)$  is the space of  $m$ th order scores for  $\rho$ ; however,  $\Gamma_m^{nuis}(\theta)$  also includes the mixed scores that have  $s$  derivatives in the direction  $\tau$  and  $m - s \geq 1$  derivatives in  $\rho$  directions. It is this strict inclusion that gives rise to higher order phenomena that do not occur in the first order theory.

**Theorem 38 :** Suppose  $\mathbb{ES}_m^{test}(\theta)$  exists in  $\mathcal{U}_m(\theta)$ . Then for  $\theta \in \Theta(\tau^\dagger)$ , (i) the set of estimation nuisance scores  $\Gamma_m^{nuis}(\theta)$  includes the set of testing nuisance scores  $\Gamma_m^{nuis, test}(\theta, \tau^\dagger)$  with equality of the sets when  $m = 1$ , (ii)  $\mathbb{IF}_{m, \tau(\cdot)}^{test}(\theta), \theta \in \Theta(\tau^\dagger)$  is standardized if and only if  $E[\mathbb{IF}_{m, \tau(\cdot)}^{test}(\theta) \mathbb{ES}_m^{test}(\theta)] = 1$  if and only if  $E[\mathbb{IF}_{m, \tau(\cdot)}^{test}(\theta) \mathbb{ES}_1^{test}(\theta)] = 1$ , (iii)

$$\left\{ \mathbb{IF}_{m, \tau(\cdot)}^{std, test} \right\} = \left\{ E_\theta [\mathbb{IF}_{m, \tau(\cdot)}^{test} \mathbb{ES}_1^{test}(\theta)]^{-1} \mathbb{IF}_{m, \tau(\cdot)}^{test}; \mathbb{IF}_{m, \tau(\cdot)}^{test} \in \{\mathbb{IF}_{m, \tau(\cdot)}^{test}\} \right\},$$

(iv) the set  $\{\mathbb{IF}_{m, \tau(\cdot)}(\theta)\}$  of all  $m$ th order estimation influence functions is contained in  $\left\{ \mathbb{IF}_{m, \tau(\cdot)}^{std, test} \right\}$  with equality of the sets when  $m = 1$ , (v)

$$\Pi_\theta \left[ \mathbb{IF}_{m, \tau(\cdot)}^{std, test}(\theta) | \Gamma_m^{test}(\theta, \tau^\dagger) \right] = \{var[\mathbb{ES}_m^{test}(\theta)]\}^{-1} \mathbb{ES}_m^{test}(\theta),$$

(vi)  $\{var_\theta[\mathbb{ES}_m^{test}(\theta)]\}^{-1} \mathbb{ES}_m^{test}(\theta) \in \left\{ \mathbb{IF}_{m, \tau(\cdot)}^{std, test} \right\}$  and has the minimum variance  $\{var_\theta[\mathbb{ES}_m^{test}(\theta)]\}^{-1}$  among members of  $\left\{ \mathbb{IF}_{m, \tau(\cdot)}^{std, test} \right\}$ . In particular  $\{var_\theta[\mathbb{ES}_m^{test}(\theta)]\}^{-1} \leq var_\theta[\mathbb{IF}_{m, \tau(\cdot)}^{eff}(\theta)]$

with equality when  $m = 1$ , (vii) Given  $\mathbb{IF}_{m,\tau(\cdot)}^{test}(\cdot) \in \{\mathbb{IF}_{m,\tau(\cdot)}^{test}(\cdot)\}$ , any smooth submodel  $\tilde{\theta}(\zeta)$  with range containing  $\theta$  and contained in  $\Theta(\tau^\dagger)$ , and  $s \leq m$ , we have

$$\partial^s E_\theta \left[ \mathbb{IF}_{m,\tau(\cdot)}^{test} \left( \tilde{\theta}(\zeta) \right) \right] / \partial \zeta_{l_1} \dots \partial \zeta_{l_s} |_{\zeta = \tilde{\theta}^{-1}\{\theta\}} = 0.$$

Thus, if  $E_\theta [\mathbb{IF}_{m,\tau(\cdot)}^{test}(\theta^*)]$  is Fréchet differentiable w.r.t.  $\theta^*$  to order  $m+1$  for a norm  $\|\cdot\|$ ,  $E_\theta [\mathbb{IF}_{m,\tau(\cdot)}^{test}(\theta + \delta\theta)] = O(\|\delta\theta\|^{m+1})$  for  $\theta$  and  $\theta + \delta\theta$  in an open neighborhood contained in  $\Theta(\tau^\dagger)$ , since the Taylor expansion of  $E_\theta [\mathbb{IF}_{m,\tau(\cdot)}^{test}(\theta^*)]$  around  $\theta$  through order  $m$  is identically zero.

The proof of the Theorem will use the following two lemmas:

**Lemma 39** :For any  $\mathbb{IF}_{m,\tau(\cdot)}^{test}(\theta)$ ,  $\theta \in \Theta(\tau^\dagger)$

$$E_\theta [\mathbb{IF}_{m,\tau(\cdot)}^{test}(\theta) \mathbb{ES}_1^{test}(\theta)] = E_\theta [\mathbb{IF}_{m,\tau(\cdot)}^{test}(\theta) \mathbb{ES}_m^{test}(\theta)]$$

**Proof.**

$$\begin{aligned} & E_\theta [\mathbb{IF}_{m,\tau(\cdot)}^{test} \mathbb{ES}_m^{test}(\theta)] \\ &= E_\theta [\mathbb{IF}_{m,\tau(\cdot)}^{test} \Pi_\theta [\mathbb{ES}_1^{test}(\theta) | \Gamma_m^{nuis,test,\perp}(\theta, \tau^\dagger)]] = E_\theta [\mathbb{IF}_{m,\tau(\cdot)}^{test} \mathbb{ES}_1^{test}(\theta)], \end{aligned}$$

where the last equality holds by  $\mathbb{IF}_{m,\tau(\cdot)}^{test} \in \Gamma_m^{nuis,test,\perp}(\theta, \tau^\dagger)$  ■

**Lemma 40** For any  $\mathbb{IF}_{m,\tau(\cdot)}^{test}(\theta)$ ,  $\theta \in \Theta(\tau^\dagger)$ ,

$$\begin{aligned} & \Pi_\theta [\mathbb{IF}_{m,\tau(\cdot)}^{test}(\theta) | \Gamma_m^{test}(\theta, \tau^\dagger)] \\ &= E [\mathbb{IF}_{m,\tau(\cdot)}^{test}(\theta) \mathbb{ES}_m^{test}(\theta)] \{var [\mathbb{ES}_m^{test}(\theta)]\}^{-1} \mathbb{ES}_m^{test}(\theta) \\ &= E [\mathbb{IF}_{m,\tau(\cdot)}^{test}(\theta) \mathbb{ES}_1^{test}(\theta)] \{var [\mathbb{ES}_m^{test}(\theta)]\}^{-1} \mathbb{ES}_m^{test}(\theta) \end{aligned}$$

**Proof.**  $\Gamma_m^{test}(\theta, \tau^\dagger) = \{c\mathbb{ES}_m^{test}(\theta); c \in R^1\} \oplus \Gamma_m^{nuis, test}(\theta, \tau^\dagger)$ . Thus, by  $\mathbb{IF}_{m, \tau(\cdot)}^{test}(\theta) \in \Gamma_m^{nuis, test, \perp}(\theta, \tau^\dagger)$ ,

$$\begin{aligned} \Pi_\theta [\mathbb{IF}_{m, \tau(\cdot)}^{test}(\theta) | \Gamma_m^{test}(\theta, \tau^\dagger)] &= \Pi_\theta [\mathbb{IF}_{m, \tau(\cdot)}^{test}(\theta) | \{c\mathbb{ES}_m^{test}(\theta); c \in R^1\}] \\ &= E [\mathbb{IF}_{m, \tau(\cdot)}^{test}(\theta) \mathbb{ES}_m^{test}(\theta)] \{var [\mathbb{ES}_m^{test}(\theta)]\}^{-1} \mathbb{ES}_m^{test}(\theta). \end{aligned}$$

Now apply Lemma 39. ■

**Proof.** (Theorem 38) (i) is immediate from the definitions. (ii) and (iii) follow from

$$\begin{aligned} E [\mathbb{IF}_{m, \tau(\cdot)}^{test}(\theta) \mathbb{ES}_m^{test}(\theta)] &= 1 \Leftrightarrow E [\mathbb{IF}_{m, \tau(\cdot)}^{test}(\theta) \mathbb{ES}_1^{test}(\theta)] = 1 \\ &\Leftrightarrow E_\theta [\mathbb{IF}_{m, \tau(\cdot)}^{test} \mathbb{IF}_{1, \tau(\cdot)}^{eff}(\theta)] = var_\theta [\mathbb{IF}_{1, \tau(\cdot)}^{eff}(\theta)], \end{aligned}$$

where we have used Lemma 39. For (iv), note  $\{\mathbb{IF}_{m, \tau(\cdot)}(\theta)\} \subset \{\mathbb{IF}_{m, \tau(\cdot)}^{test}\}$  follows from the fact that every smooth submodel through  $\theta$  in model  $\mathcal{M}(\Theta(\tau^\dagger))$  is a smooth submodel through  $\theta$  in model  $\mathcal{M}(\Theta)$ . Thus it remains to prove that  $\mathbb{IF}_{m, \tau(\cdot)}(\theta)$  is standardized. But, by Part 4 of Theorem 3,  $E_\theta [\mathbb{IF}_{m, \tau(\cdot)}(\theta) \mathbb{IF}_{1, \tau(\cdot)}^{eff}(\theta)] = var_\theta [\mathbb{IF}_{1, \tau(\cdot)}^{eff}(\theta)]$ . (v) follows at once from Lemma 39 and Part (ii). For (vi), note that  $\{var_\theta [\mathbb{ES}_m^{test}(\theta)]\}^{-1} \mathbb{ES}_m^{test}(\theta) \in \{\mathbb{IF}_{m, \tau(\cdot)}^{std, test}\}$  by definition. Thus

$$var_\theta \left\{ E_\theta [\mathbb{IF}_{m, \tau(\cdot)}^{test} \mathbb{ES}_m^{test}(\theta)]^{-1} \mathbb{IF}_{m, \tau(\cdot)}^{test} \right\} \geq \{var_\theta [\mathbb{ES}_m^{test}(\theta)]\}^{-1}$$

follows from (v). The result then follows from part (iii). Part (vii) is proved analogously to Theorem 2 except now all scores lie in  $\Gamma_m^{nuis}(\theta)$  by range  $\tilde{\theta}(\zeta)$  in  $\Theta(\tau^\dagger)$ .

■

In the case of (locally) nonparametric models, we can explicitly characterize  $\Gamma_m^{test,\perp}(\theta, \tau^\dagger)$ . Let  $\{\mathbb{U}_{j,j}^{test,\perp}(\theta, \tau^\dagger)\}$  be the set of all  $\mathbb{U}_{j,j}^{test,\perp}(\theta, \tau^\dagger) = \mathbb{V}\left[U_{j,j}^{test,\perp}(\theta, \tau^\dagger)\right]$  with the  $U_{j,j}^{test,\perp}(\theta, \tau^\dagger) = \sum_{l=1}^{\infty} c_l IF_{1,\tau(\cdot),i_1}^{eff}(\theta) \prod_{s=2}^j h_{l,s}(O_{i_s}; \theta) \in \mathcal{U}_j(\theta)$ , indexed by constants  $c_l \in R^1$ , and functions  $h_{l,s}(O_{i_s}; \theta)$  satisfying  $E_\theta[h_{l,s}(O_{i_s}; \theta)] = 0$ . We remark that the subset of  $\mathcal{U}_j(\theta)$  comprised of all  $j$ th order degenerate U-statistics can be written  $\left\{ \mathbb{V}\left[\sum_{l=1}^{\infty} \prod_{s=1}^j h_{l,s}(O_{i_s}; \theta)\right] \right\}$ . Thus  $\{\mathbb{U}_{j,j}^{test,\perp}(\theta, \tau^\dagger)\}$  simply restricts one of the functions  $h_{l,s}(O; \theta)$  to be  $c_l IF_{1,\tau(\cdot)}^{eff}$ .

**Theorem 41** *If the model  $\mathcal{M}(\Theta)$  is (locally) nonparametric, then  $\Gamma_m^{test,\perp}(\theta, \tau^\dagger) = \left\{ \sum_{j=2}^m \mathbb{U}_{j,j}^{test,\perp}(\theta, \tau^\dagger); \mathbb{U}_{j,j}^{test,\perp}(\theta, \tau^\dagger) \in \left\{ \mathbb{U}_{j,j}^{test,\perp}(\theta, \tau^\dagger) \right\} \right\}$ .*

**Proof.** Since the model is locally nonparametric  $\Gamma_m^{test}(\theta, \tau^\dagger)$  includes the set of all mean zero first order  $U$ -statistics  $\mathcal{U}_1(\theta)$  and thus any element of  $\Gamma_m^{test,\perp}(\theta, \tau^\dagger)$  must be a sum of degenerate  $U$ -statistics of orders 2 through  $m$ . We continue by induction. First we prove the theorem for  $m = 2$ . Now,  $\Gamma_2^{test}(\theta, \tau^\dagger) = \mathcal{U}_1(\theta) + \mathcal{U}_{2,2}^{nuis,test}(\theta)$  where  $\mathcal{U}_{2,2}^{nuis,test}$  is the closed linear span of the 2nd order degenerate part  $\sum_{s \neq j} S_{l_1,j} S_{l_2,s}$  of 2nd order scores  $\tilde{S}_{2,\bar{l}_2} = \sum_j S_{l_1 l_2, j} + \sum_{s \neq j} S_{l_1, j} S_{l_2, s}$  in model  $\mathcal{M}(\Theta(\tau^\dagger))$ , where  $\sum_{s \neq j} S_{l_1, j} S_{l_2, s}$  is a sum of products  $S_{l_1, j} S_{l_2, s}$  of first order scores in model  $\mathcal{M}(\Theta(\tau^\dagger))$  for two different subjects. By model  $\mathcal{M}(\Theta)$  being (locally) nonparametric, the set of first order scores in model  $\mathcal{M}(\Theta(\tau^\dagger))$  is precisely the set of random variables  $\Gamma_1^{nuis,test}(\theta, \tau^\dagger)$  orthogonal to  $IF_{1,\tau(\cdot)}^{eff}(\theta)$ . But the set of degenerate  $U$ -statistics of order 2 orthogonal to the product of two scores in  $\Gamma_1^{nuis,test}(\theta, \tau^\dagger)$  is

clearly  $\left\{ \mathbb{U}_{2,2}^{test,\perp}(\theta, \tau^\dagger) \right\}$ . Suppose now the theorem is true for  $m, m \geq 2$ , we show it is true for  $m + 1$ . By  $\mathcal{M}(\Theta)$  (locally) nonparametric and the induction assumption,  $\Gamma_{m+1}^{test}(\theta, \tau^\dagger) = \Gamma_m^{test}(\theta, \tau^\dagger) + \mathcal{U}_{m+1,m+1}^{test}(\theta)$  where  $\mathcal{U}_{m+1,m+1}^{nuis,test}(\theta)$  is the closed linear span of the sum of products of first order scores in model  $\mathcal{M}(\Theta(\tau^\dagger))$  for  $m + 1$  different subjects. But  $\left\{ \mathbb{U}_{m+1,m+1}^{test,\perp}(\theta, \tau^\dagger) \right\}$  is the set of set of degenerate  $U - statistics$  of order  $m + 1$  orthogonal to  $\mathcal{U}_{m+1,m+1}^{nuis,test}(\theta)$ . ■

### 6.3 Implicitly defined Functionals and Testing Influence Functions:

In the following theorem we show that estimation influence functions  $\mathbb{IF}_{m,\psi(\tau,\cdot)}(\theta)$  for the parameter  $\psi(\tau, \cdot)$  evaluated at the solution  $\tau(\theta)$  to  $0 = \psi(\tau, \theta)$  is contained in the set  $\left\{ \mathbb{IF}_{m,\tau(\cdot)}^{test}(\theta) \right\}$  of testing influence functions for  $\tau(\theta)$ . We also derive the estimation influence functions  $\mathbb{IF}_{m,\tau(\cdot)}(\theta) = \sum_{s=1}^m \mathbb{IF}_{s,s,\tau(\cdot)}(\theta)$  for  $\tau(\theta)$  in terms of the estimation influence functions  $\mathbb{IF}_{m,\psi(\tau,\cdot)}(\theta)$  for  $\psi(\tau, \cdot)$  and their derivatives with respect to  $\tau$ .

**Theorem 42** *Let  $\tau(\theta)$  be the assumed unique functional defined by  $0 = \psi(\tau(\theta), \theta), \theta \in \Theta$ . Then, for  $\theta \in \Theta(\tau^\dagger)$ , whenever  $\mathbb{IF}_{m,\psi(\tau^\dagger,\cdot)}(\theta)$  and  $\mathbb{IF}_{m,\tau(\cdot)}(\theta)$  exist, (i)  $\mathbb{IF}_{m,\psi(\tau^\dagger,\cdot)}(\theta) \in \left\{ \mathbb{IF}_{m,\tau(\cdot)}^{test}(\theta) \right\}$ , (ii)  $\mathbb{IF}_{1,\tau(\cdot)}(\theta) = -\psi_\tau^{-1} \mathbb{IF}_{1,\psi(\tau^\dagger,\cdot)}(\theta) \in \left\{ \mathbb{IF}_{1,\tau(\cdot)}^{std,test}(\theta) \right\}$  where  $\psi_\tau \equiv \partial \psi(\tau, \theta) / \partial \tau|_{\tau=\tau^\dagger}$ , (iii)  $\mathbb{IF}_{m,m,\tau(\cdot)}(\theta) = -\psi_\tau^{-1} \left\{ \mathbb{IF}_{m,m,\psi(\tau^\dagger,\cdot)}(\theta) + \mathbb{Q}_{m,m}(\theta) \right\}$ , where  $\mathbb{Q}_{m,m}(\theta) \equiv$*

$\mathbb{Q}_{m,m,\tau(\cdot)}(\theta) = \mathbb{V}\{Q_{m,m}(\theta)\} \in \{\mathbb{U}_{m,m}^{test,\perp}(\theta, \tau^\dagger)\}$ . For  $m = 2$ ,

$$Q_{2,2}(\theta) = \frac{1}{2} \psi_{\setminus \tau^2} IF_{1,\tau(\cdot),i_1}(\theta) IF_{1,\tau(\cdot),i_2}(\theta) + \frac{1}{2} \left[ \begin{aligned} & \left( \begin{aligned} & \frac{\partial IF_{1,\psi(\tau^\dagger,\cdot),i_1}(\theta)}{\partial \tau} \\ & - E_\theta \left[ \frac{\partial IF_{1,\psi(\tau^\dagger,\cdot),i_1}(\theta)}{\partial \tau} \right] \end{aligned} \right) IF_{1,\tau(\cdot),i_2}(\theta) \\ & + \left( \begin{aligned} & \frac{\partial IF_{1,\psi(\tau^\dagger,\cdot),i_2}(\theta)}{\partial \tau} \\ & - E_\theta \left[ \frac{\partial IF_{1,\psi(\tau^\dagger,\cdot),i_2}(\theta)}{\partial \tau} \right] \end{aligned} \right) IF_{1,\tau,i_1}(\theta) \end{aligned} \right] \quad (52)$$

where  $\frac{\partial IF_{1,\psi(\tau^\dagger,\cdot),i_1}(\theta)}{\partial \tau} = \partial IF_{1,\psi(\tau,\cdot),i_1}(\theta) / \partial \tau|_{\tau=\tau^\dagger}$ .  $Q_{m,m}(\theta)$  is given in the appendix as well as the general formula.

**Proof.** (i) For  $r \leq m$ , consider any suitably smooth  $r$  dimensional parametric sub-model  $\tilde{\theta}(\zeta)$  with range containing  $\theta$  and contained in  $\Theta(\tau^\dagger)$ . Let  $\tilde{\mathbb{S}}_{s\setminus \bar{l}_s}(\theta)$  be any associated  $sth$ -order score  $s \leq m$ . By definition of  $\tau(\theta)$ ,  $\psi(\tau(\theta(\zeta)), \theta(\zeta)) = 0$ . Hence,  $0 = \partial^s \psi(\tau(\theta(\zeta)), \theta(\zeta)) / \partial \zeta_{l_1} \dots \partial \zeta_{l_s}|_{\zeta=\tilde{\theta}^{-1}(\theta)}$ . Now we expand the RHS using the chain rule and note that the only non-zero term is the term  $\psi_{\setminus \bar{l}_s}(\tau^\dagger, \theta)$  in which all  $s$ -derivatives are taken with respect to the second  $\theta(\zeta)$  in  $\psi(\tau(\theta(\zeta)), \theta(\zeta))$ ; all other terms include derivatives of  $\tau(\theta(\zeta))$ , which are zero by range  $\tilde{\theta}(\zeta) \subset \Theta(\tau^\dagger)$ . Further  $\psi_{\setminus \bar{l}_s}(\tau^\dagger, \theta) = E_\theta [\mathbb{IF}_{m,\psi(\tau^\dagger,\cdot)}(\theta) \tilde{\mathbb{S}}_{s\setminus \bar{l}_s}(\theta)]$  by the definition of the estimation influence function  $\mathbb{IF}_{m,\psi(\tau^\dagger,\cdot)}(\theta)$ . We conclude that  $\mathbb{IF}_{m,\psi(\tau^\dagger,\cdot)}(\theta)$  is in  $\Gamma_m^{nuis,test}(\theta, \tau^\dagger)^\perp$ . (ii)  $\mathbb{IF}_{1,\tau(\cdot)} = -\psi_\tau^{-1} \mathbb{IF}_{1,\psi(\tau^\dagger,\cdot)}$  is straightforward. That  $\mathbb{IF}_{1,\tau(\cdot)}$  is contained in  $\{\mathbb{IF}_{1,\tau(\cdot)}^{std,test}\}$  follows by Part (iv) of Theorem 38. (iii) See appendix for proof. ■

#### 6.4 "Inefficiency" of the Efficient Score

We now provide an example to show that, contrary to what one might expect based on Part (vi) of Theorem 38, inference concerning  $\tau(\theta)$  may be more efficient when based on an 'inefficient' member of the set  $\{\mathbb{IF}_{m,\tau(\cdot)}^{test}(\theta)\}$  such as  $\mathbb{IF}_{m,\psi(\tau^\dagger,\cdot)}(\theta)$  than when based on the efficient score  $\mathbb{ES}_{m,\tau(\cdot)}^{test}(\theta)$ . Without loss of generality, it is sufficient to consider the case  $m = 2$ . In the following example it is  $\tilde{\tau}_k(\theta)$  and  $\tilde{\psi}_k(\tau^\dagger, \theta)$  that play the role of  $\tau(\theta)$  and  $\psi(\tau^\dagger, \theta)$  in the preceding theorem, because  $\tilde{\tau}_k(\theta)$  and  $\tilde{\psi}_k(\tau^\dagger, \theta)$  have, but  $\tau(\theta)$  and  $\psi(\tau^\dagger, \theta)$  do not have, higher order estimation and testing influence functions.

**Example 1c (continued):** In this example, with  $Y^*(\tau) \equiv Y^* - \tau A$ ,  $A$  and  $Y^*$  binary,

$$\psi(\tau, \theta) = E_\theta[\{Y^*(\tau) - E_\theta(Y^*(\tau)|X)\}\{A - E_\theta(A|X)\}]$$

and  $\tau(\theta)$  satisfies  $\psi(\tau(\theta), \theta) = 0$ . Let  $\tilde{\tau}_k(\theta)$  satisfy  $\tilde{\psi}_k(\tilde{\tau}_k(\theta), \theta) = 0$  where  $\tilde{\psi}_k(\tau, \theta) = E_\theta[Y^*(\tau)A] - E_\theta\{\Pi_\theta[B(\tau)|\bar{Z}_k]\Pi_\theta[P|\bar{Z}_k]\}$  is defined in Section 3.1 with  $\tau$  a real-valued index and  $B(\tau) = b(X, \tau) = E_\theta(Y^*(\tau)|X)$ . Note  $\tilde{\psi}_{k,\tau}(\tau, \theta) \equiv \partial\tilde{\psi}_k(\tau, \theta)/\partial\tau = -\{E_\theta[A^2] - E_\theta[\{\Pi_\theta[P|\bar{Z}_k]\}^2]\}$ ,  $\psi_\tau(\tau, \theta) = -E_\theta[var_\theta(A|X)]$ ,  $\tilde{\psi}_{k,\tau^2}(\tau, \theta) = \psi_{\tau^2}(\tau, \theta) = 0$ . Below we freely use results of Theorems 18, 20, and 23. We suppose that  $0 < \sigma < var_\theta(A|X)$  and  $E_\theta[A^2] < c$  for some  $(\sigma, c)$ ,  $\beta = \frac{\beta_p + \beta_b}{2} < 1/4$ . Choose  $k = k_{opt}(2)n^{2\sigma} = n^{\frac{2}{1+4\beta}+2\sigma}$ ,  $\sigma > 0$  so the truncation bias of  $\hat{\psi}_{2,k}(\tau) \equiv \psi_{2,k}(\tau, \hat{\theta})$  is  $O_p\left(n^{-\frac{4\beta}{4\beta+d}}\right)$  and  $n^{-\frac{4\beta}{4\beta+d}} \ll var_\theta[\hat{\psi}_{2,k}(\tau)] \asymp k/n^2 = n^{-2(\frac{4\beta}{4\beta+d}+\sigma)}$ . We assume the



given  $(\beta_g, \beta_b, \beta_p)$  are such that the order  $O_p \left[ n^{-\left(\frac{\beta_g}{2\beta_g+d} + \frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p}\right)} \right]$  of the estimation bias of  $\widehat{\psi}_{2,k}(\tau)$  is  $O_p \left( n^{-\frac{4\beta}{4\beta+d}} \right)$ . Then  $\left| \widehat{\psi}_{2,k}(\tau) - \widetilde{\psi}_k(\tau, \theta) \right|$  and  $\left| \widehat{\psi}_{2,k}(\tau) - \psi(\tau, \theta) \right|$  are  $O_p \left( n^{-\frac{4\beta}{4\beta+d} + \sigma} \right)$  which just exceeds the minimax rate  $O_p \left( n^{-\frac{4\beta}{4\beta+d}} \right)$  for  $\sigma$  very small.

Our goal is to compare the coverage and length of confidence intervals for  $\widetilde{\tau}_k(\theta)$  and  $\tau(\theta)$  based on

$$\begin{aligned} C_{1-\alpha, \widetilde{\psi}_k(\tau)} &\equiv \left\{ \tau; -z_{1-\alpha/2} < \frac{\psi_{2,k}(\tau, \widehat{\theta})}{\mathbb{W}_{2, \widetilde{\psi}_k(\tau)}(\widehat{\theta})} < z_{1-\alpha/2} \right\}, \\ C_{1-\alpha, 2, \widetilde{\tau}_k} &\equiv \left\{ \tau; -z_{1-\alpha/2} < \frac{\tau_{2,k}(\widehat{\theta}) - \tau}{\mathbb{W}_{2, \widetilde{\tau}_k}(\widehat{\theta})} < z_{1-\alpha/2} \right\}, \\ C_{1-\alpha, 2, ES} &\equiv \left\{ \tau; -z_{1-\alpha/2} < \frac{\mathbb{ES}_{2, \widetilde{\tau}_k}^{test}(\widehat{\theta}(\tau))}{\mathbb{W}_{2, \widetilde{\tau}_k}^{ES}(\widehat{\theta}(\tau))} < z_{1-\alpha/2} \right\}, \end{aligned}$$

where  $\mathbb{W}_{2, \widetilde{\psi}_k(\tau)}(\widehat{\theta})$ ,  $\mathbb{W}_{2, \widetilde{\tau}_k}(\widehat{\theta})$ ,  $\mathbb{W}_{2, \widetilde{\tau}_k}^{ES}(\widehat{\theta}(\tau))$  are appropriate variance estimators,  $\widehat{\theta}$  is our usual split sample initial estimator, and  $\widehat{\theta}(\tau^\dagger)$  is an initial split sample estimator depending on  $\tau^\dagger$  that satisfies  $\widetilde{\psi}_k(\tau, \widehat{\theta}(\tau^\dagger)) = 0$  if  $\tau = \tau^\dagger$ , i.e.,  $\tau[\widehat{\theta}(\tau^\dagger)] = \tau^\dagger$ . We assume that if  $\tau(\theta) = \tau^\dagger$  then the convergence rate under  $\theta$  of our estimator of  $b(X, \tau^*)$  for any  $\tau^*$  remains  $n^{-\frac{\beta_b}{d+2\beta_b}}$ . Now the assumption  $0 < \sigma < E_\theta[\text{var}_\theta(A|X)]$ ,  $E_\theta[A^2] < c$  implies  $|\widetilde{\tau}_k(\theta) - \tau(\theta)| / \left| \widetilde{\psi}_k(\tau, \theta) - \psi(\tau, \theta) \right|$  is uniformly bounded away from zero and infinity. It then follows from earlier results on  $\widehat{\psi}_{2,k}(\tau)$ , the assumption  $0 < \sigma < E_\theta[\text{var}_\theta(A|X)]$ ,  $E_\theta[A^2] < c$ , and Theorem 42 that  $C_{1-\alpha, \widetilde{\psi}_k(\tau)}$  is a uniform asymptotic  $1-\alpha$  confidence interval for both  $\tau(\theta)$  and  $\widetilde{\tau}_k(\theta)$  of length  $O_p \left( n^{-\frac{4\beta}{4\beta+d} + \sigma} \right)$ .

The next theorem gives explicit formulae for  $\psi_{2,k}(\tau, \hat{\theta})$ ,  $\mathbb{ES}_{2,\tilde{\tau}_k}^{test}(\hat{\theta}(\tau))$ , and  $\tau_{2,k}(\hat{\theta})$ .

Using these formulae we calculate the biases and variances necessary to compare the coverage of the three intervals.

This comparison requires each of our three candidate procedures to be on the same scale. Therefore we used standardized versions of the relevant statistics.

**Theorem 43** *Suppose the assumptions described in the preceding example hold. Then*

(i)

$$\begin{aligned}\psi_{2,k}(\tau, \hat{\theta}) &= \tilde{\psi}_{k,\tau}(\tau, \hat{\theta}) + \mathbb{IF}_{2,\tilde{\psi}_k(\tau,\cdot)}(\hat{\theta}) \\ &= \tilde{\psi}_{k,\tau}(\tau, \hat{\theta}) + \mathbb{V}\left[\left(Y^*(\tau) - \hat{b}(X, \tau)\right)\{A - \hat{p}(X)\}\right] \\ &\quad + \mathbb{V}\left[\left\{\left[Y^*(\tau) - \hat{b}(X, \tau)\right]\bar{Z}_k^T\right\}_{i_1}\left\{\bar{Z}_k[A - \hat{p}(X)]\right\}_{i_2}\right]\end{aligned}$$

where  $\hat{b}(X, \tau) = \hat{B}(\tau) = E_{\hat{\theta}}(Y^*(\tau) | X)$ ,  $\hat{p}(X) = \hat{P} = E_{\hat{\theta}}(A | X)$ ;

(ii): Let  $\hat{\epsilon}$  denote  $Y - \hat{b}(X)$ , and  $\hat{\Delta}$  denote  $A - \hat{p}(X)$ . Thus,

$$\begin{aligned}&\mathbb{ES}_{2,\tilde{\tau}_k}^{test}(\hat{\theta}(\tau^\dagger)) \\ &= v(\hat{\theta}(\tau^\dagger))\left\{var_{\hat{\theta}(\tau^\dagger)}\left[\mathbb{IF}_{2,\tilde{\psi}_k(\tau^\dagger,\cdot)}(\hat{\theta}(\tau^\dagger))\right] - var\left[\mathbb{U}_{2,2,\tilde{\tau}_k(\cdot)}^{*,test,\perp}(\hat{\theta}(\tau^\dagger), \tau^\dagger)\right]\right\}^{-1} \\ &\times \left\{\mathbb{IF}_{2,\tilde{\psi}_k(\tau^\dagger,\cdot)}(\hat{\theta}(\tau^\dagger)) - \mathbb{U}_{2,2,\tilde{\tau}_k(\cdot)}^{*,test,\perp}(\hat{\theta}(\tau^\dagger), \tau^\dagger)\right\}\end{aligned}$$

where

$$\begin{aligned}
& U_{2,2,\tilde{\tau}_k(\cdot),ij}^{*,test,\perp} \left( \hat{\theta}(\tau^\dagger), \tau^\dagger \right) \\
&= \left( E_{\hat{\theta}} \left[ \hat{\epsilon}_i^2 \hat{\Delta}_i^2 \right] \right)^{-1} \times \hat{\epsilon}_i \hat{\Delta}_i \\
& \left( - \left\{ \begin{aligned} & \left( E_{\hat{\theta}} \left[ \hat{\epsilon}_i^2 \hat{\Delta}_i^2 \right] \right)^{-1} E_{\hat{\theta}} \left[ \hat{\epsilon} \hat{\Delta}^2 \overline{Z}_k^T \right] \\ & \times E_{\hat{\theta}} \left[ \hat{\epsilon}^2 \hat{\Delta} \overline{Z}_k^T \right] \hat{\epsilon}_j \hat{\Delta}_j \\ & + E_{\hat{\theta}} \left[ \hat{\epsilon}_i^2 \hat{\Delta}_i \overline{Z}_{k,i}^T \right] \overline{Z}_{k,j} \hat{\Delta}_j \\ & + E_{\hat{\theta}} \left[ \hat{\epsilon}_i \hat{\Delta}_i^2 \overline{Z}_{k,i}^T \right] \overline{Z}_{k,j} \hat{\epsilon}_j \end{aligned} \right\} \right)
\end{aligned}$$

and

$$v(\theta) = E_\theta [var_\theta(A|X)]$$

Also,

$$\begin{aligned}
& var_{\hat{\theta}(\tau^\dagger)} \left\{ \mathbb{E}S_{2,\tilde{\tau}_k(\cdot)}^{test} \left( \hat{\theta}(\tau^\dagger) \right) \right\}^{-1} \mathbb{E}S_{2,\tilde{\tau}_k(\cdot)}^{test} \left( \hat{\theta}(\tau^\dagger) \right) \\
&= v \left( \hat{\theta}(\tau^\dagger) \right)^{-1} \left\{ \mathbb{I}\mathbb{F}_{2,\tilde{\psi}_k(\tau^\dagger,\cdot)} \left( \hat{\theta}(\tau^\dagger) \right) - \mathbb{U}_{2,2,\tilde{\tau}_k(\cdot)}^{*,test,\perp} \left( \hat{\theta}(\tau^\dagger), \tau^\dagger \right) \right\} \quad (53)
\end{aligned}$$

(iii)

$$\tau_{2,k} \left( \hat{\theta} \right) \equiv \tilde{\tau}_k \left( \hat{\theta} \right) + \mathbb{I}\mathbb{F}_{1,\tilde{\tau}_k(\cdot)} \left( \hat{\theta} \right) + \mathbb{I}\mathbb{F}_{2,2,\tilde{\tau}_k(\cdot)} \left( \hat{\theta} \right),$$

where

$$\mathbb{I}\mathbb{F}_{1,\tilde{\tau}_k(\cdot)} \left( \hat{\theta} \right) = \mathbb{I}\mathbb{F}_{1,\tilde{\tau}_k(\cdot)} \left( \hat{\theta} \right) = \mathbb{V} \left\{ v \left( \hat{\theta} \right)^{-1} \left[ \left\{ Y - \hat{b}(X) \right\} \left\{ A - \hat{p}(X) \right\} \right] \right\}$$

with  $Y = Y^* \left( \tau \left( \hat{\theta} \right) \right)$ ,  $\hat{b}(X) = \hat{b} \left( X, \tau \left( \hat{\theta} \right) \right)$ ,

$$\mathbb{IF}_{2,2,\tilde{\tau}_k(\cdot)}(\hat{\theta}) = v(\hat{\theta})^{-1} \left[ \mathbb{IF}_{2,2,\tilde{\psi}_k(\tau(\hat{\theta}),\cdot)}(\hat{\theta}) + \mathbb{Q}_{2,2,\tilde{\tau}_k(\cdot)}(\hat{\theta}) \right] \text{ where}$$

$$\begin{aligned} & \mathbb{Q}_{2,2,\tilde{\tau}_k(\cdot),\tilde{b}_2}(\hat{\theta}) \\ &= -\frac{1}{2}v(\hat{\theta})^{-1} \left[ \begin{aligned} & \left[ \{A - \hat{p}(X)\}_{i_1}^2 - v(\hat{\theta}) \right] \left[ \{Y - \hat{b}(X)\} \{A - \hat{p}(X)\} \right]_{i_2} + \\ & \left[ \{A - \hat{p}(X)\}_{i_2}^2 - v(\hat{\theta}) \right] \left[ \{Y - \hat{b}(X)\} \{A - \hat{p}(X)\} \right]_{i_1} \end{aligned} \right] \end{aligned}$$

**Proof.** The proof of (i) was given earlier. The proofs of (ii) and (iii) are in the appendix. ■

**Theorem 44 .** Suppose  $\tilde{\tau}_k(\theta) = \tau^\dagger$  and the assumptions of the preceding theorem hold. Then

$$\begin{aligned} & (i) \text{ var}_\theta \left[ \mathbb{U}_{2,2,\tilde{\tau}_k(\cdot)}^{*,test,\perp}(\hat{\theta}(\tau^\dagger), \tau^\dagger) \right] = o\left(\frac{1}{n}\right), \\ & \text{var}_\theta \left[ v(\hat{\theta})^{-1} \psi_{2,k}(\tau, \hat{\theta}) \right] \times \left[ \text{var}_\theta \left\{ \text{var}_{\hat{\theta}(\tau^\dagger)} \left\{ \mathbb{ES}_{2,\tilde{\tau}_k(\cdot)}^{test}(\hat{\theta}(\tau^\dagger)) \right\}^{-1} \mathbb{ES}_{2,\tilde{\tau}_k(\cdot)}^{test}(\hat{\theta}(\tau^\dagger)) \right\} \right]^{-1} \\ &= 1 + o_p(1) \end{aligned}$$

(ii)

$$\begin{aligned} & \text{var}_\theta \left[ \mathbb{Q}_{2,2,\tilde{\tau}_k(\cdot)}(\hat{\theta}) \right] = o\left(\frac{1}{n}\right), \\ & \text{var}_\theta \left[ v(\hat{\theta})^{-1} \psi_{2,k}(\tau, \hat{\theta}) \right] / \text{var}_\theta \left\{ \tau_{2,k}(\hat{\theta}) - \tau^\dagger \right\} = 1 + o_p(1) \end{aligned}$$

(iii)

$$\begin{aligned} & v(\hat{\theta})^{-1} E_\theta \left[ \psi_{2,k}(\tau^\dagger, \hat{\theta}) \right] = O_p \left\{ (P - \hat{P})(B(\tau^\dagger) - \hat{B}(\tau^\dagger)) \left( \frac{g(X)}{\hat{g}(X)} - 1 \right) \right\} \\ &= O_p \left( n^{-\left( \frac{\beta_g}{2\beta_g+d} + \frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p} \right)} \right) \end{aligned}$$

(iv)

$$\begin{aligned}
& E_{\theta} \left[ \text{var}_{\widehat{\theta}(\tau^{\dagger})} \left\{ \mathbb{ES}_{2, \tilde{\tau}_k(\cdot)}^{test} \left( \widehat{\theta}(\tau^{\dagger}) \right) \right\}^{-1} \mathbb{ES}_{2, \tilde{\tau}_k}^{test} \left( \widehat{\theta}(\tau^{\dagger}) \right) \right] \\
&= O_p \left\{ \left( P - \widehat{P} \right) \left( B(\tau^{\dagger}) - \widehat{B}(\tau^{\dagger}) \right) \left[ \left( \frac{g(X)}{\widehat{g}(X)} - 1 \right) + \left( P - \widehat{P} \right) + \left( B(\tau^{\dagger}) - \widehat{B}(\tau^{\dagger}) \right) \right] \right\} \\
&= O_p \left[ \max \left\{ n^{-\left( \frac{\beta_g}{2\beta_g+d} + \frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p} \right)}, n^{-\left( \frac{\beta_b}{d+2\beta_b} + \frac{2\beta_p}{d+2\beta_p} \right)}, n^{-\left( \frac{2\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p} \right)} \right\} \right]
\end{aligned}$$

(v)

$$\begin{aligned}
E_{\theta} \left[ \tau_{2,k} \left( \widehat{\theta} \right) - \tau^{\dagger} \right] &= O_p \left\{ \begin{aligned} & \left( P - \widehat{P} \right) \left( \frac{g(X)}{\widehat{g}(X)} - 1 \right) \left( B - \widehat{B} \right) + \\ & \left( P - \widehat{P} \right)^2 \left( P - \widehat{P} \right) \left( B - \widehat{B} \right) \\ & + \left( \frac{g(X)}{\widehat{g}(X)} - 1 \right)^2 \left[ \left( P - \widehat{P} \right) + \left( \frac{g(X)}{\widehat{g}(X)} - 1 \right) + \left( B - \widehat{B} \right) \right] \end{aligned} \right\} \\
&= O_p \left\{ \max \left\{ \begin{aligned} & n^{-\left( \frac{\beta_g}{2\beta_g+d} + \frac{\beta_p}{d+2\beta_p} + \frac{\beta_b}{d+2\beta_b} \right)}, \\ & n^{-\left( \frac{2\beta_p}{d+2\beta_p} \right)} n^{-\frac{\beta_p}{d+2\beta_p}} n^{-\frac{\beta_b}{d+2\beta_b}}, \\ & n^{-\frac{2\beta_g}{2\beta_g+d}} \left\{ n^{-\frac{\beta_b}{d+2\beta_b}} + n^{-\frac{\beta_p}{d+2\beta_p}} + n^{-\frac{\beta_g}{2\beta_g+d}} \right\} \end{aligned} \right\} \right\}
\end{aligned}$$

**Proof.** The proof of part (iii) was given earlier. The remaining parts are proved in the Appendix. ■

We conclude from this theorem that the savings in variance that comes with using  $\mathbb{ES}_{2, \tilde{\tau}_k(\cdot)}^{test} \left( \widehat{\theta}(\tau^{\dagger}) \right)$  rather than  $\psi_{2,k} \left( \tau, \widehat{\theta} \right)$  is asymptotically negligible even in regard to constants. Similarly, we conclude that the difference in variance that comes with using  $\psi_{2,k} \left( \tau, \widehat{\theta} \right)$  rather than  $\mathbb{IF}_{2,2, \tilde{\tau}_k(\cdot)} \left( \widehat{\theta} \right)$  is asymptotically negligible, again even in regard to constants. Further, because  $\text{var}_{\theta} \left[ \mathbb{U}_{2,2, \tilde{\tau}_k(\cdot)}^{*,test,\perp} \left( \widehat{\theta}(\tau^{\dagger}), \tau^{\dagger} \right) \right]$  and  $\text{var}_{\theta} \left[ \mathbb{Q}_{2,2, \tilde{\tau}_k(\cdot)} \left( \widehat{\theta} \right) \right]$  are of the order of  $o\left(\frac{1}{n}\right)$  as their first order degenerate kernels are

both of order  $o_p(1)$ , and  $n^{\frac{4\beta}{4\beta+d}-\sigma} \left\{ \psi_{2,k} \left( \tau, \hat{\theta} \right) - E_{\theta} \left[ \psi_{2,k} \left( \tau, \hat{\theta} \right) \right] \right\}$  is asymptotically normal, we conclude that

$$n^{\frac{4\beta}{4\beta+d}-\sigma} \left\{ \tau_{2,k} \left( \hat{\theta} \right) - E_{\theta} \left[ \tau_{2,k} \left( \hat{\theta} \right) \right] \right\},$$

$$n^{\frac{4\beta}{4\beta+d}-\sigma} \left\{ \mathbb{ES}_{2,\tilde{\tau}_k(\cdot)}^{test} \left( \hat{\theta}(\tau^\dagger) \right) \right\}^{-1} \left[ \mathbb{ES}_{2,\tilde{\tau}_k(\cdot)}^{test} \left( \hat{\theta}(\tau^\dagger) \right) - E_{\theta} \left[ \mathbb{ES}_{2,\tilde{\tau}_k(\cdot)}^{test} \left( \hat{\theta}(\tau^\dagger) \right) \right] \right]$$

and  $n^{\frac{4\beta}{4\beta+d}-\sigma} v \left( \hat{\theta} \right)^{-1} \left\{ \psi_{2,k} \left( \tau, \hat{\theta} \right) - E_{\theta} \left[ \psi_{2,k} \left( \tau, \hat{\theta} \right) \right] \right\}$  are all asymptotically normal with the same asymptotic variance.

It then follows that a necessary condition for the intervals based on  $\psi_{2,k} \left( \tau^\dagger, \hat{\theta} \right)$ ,  $\mathbb{ES}_{2,\tilde{\tau}_k(\cdot)}^{test} \left( \hat{\theta}(\tau) \right)$ , and  $\tau_{2,k} \left( \hat{\theta} \right) - \tau$  to cover  $\tilde{\tau}_k(\theta) = \tau^\dagger$  at the nominal  $1 - \alpha$  level as  $n \rightarrow \infty$  is that

$$v \left( \hat{\theta} \right)^{-1} E_{\theta} \left[ \psi_{2,k} \left( \tau^\dagger, \hat{\theta} \right) \right],$$

$$var_{\hat{\theta}(\tau^\dagger)} \left\{ \mathbb{ES}_{2,\tilde{\tau}_k(\cdot)}^{test} \left( \hat{\theta}(\tau^\dagger) \right) \right\}^{-1} E_{\theta} \left[ \mathbb{ES}_{2,\tilde{\tau}_k}^{test} \left( \hat{\theta}(\tau^\dagger) \right) \right]$$

and  $E_{\theta} \left[ \tau_{2,k} \left( \hat{\theta} \right) - \tau^\dagger \right]$  are  $O_p \left( n^{-\frac{4\beta}{4\beta+d}+\sigma} \right)$ .

Now we know under the assumptions of theorem 44 that this necessary condition holds for  $v \left( \hat{\theta} \right)^{-1} E_{\theta} \left[ \psi_{2,k} \left( \tau^\dagger, \hat{\theta} \right) \right]$  since  $v \left( \hat{\theta} \right)$  is bounded away from zero and one and, by assumption,  $n^{-\left(\frac{\beta_g}{2\beta_g+d} + \frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p}\right)} = O_p \left( n^{-\frac{4\beta}{4\beta+d}} \right)$ . However this necessary condition need not hold for either

$$var_{\hat{\theta}(\tau^\dagger)} \left\{ \mathbb{ES}_{2,\tilde{\tau}_k(\cdot)}^{test} \left( \hat{\theta}(\tau^\dagger) \right) \right\}^{-1} E_{\theta} \left[ \mathbb{ES}_{2,\tilde{\tau}_k}^{test} \left( \hat{\theta}(\tau^\dagger) \right) \right]$$

or  $E_{\theta} \left[ \tau_{2,k} \left( \hat{\theta} \right) - \tau^\dagger \right]$ . For example, consider the following specification consistent with our assumptions:  $\beta_p/d = 0$ ,  $\beta_b/d = \beta_g/d = 1/4$ . Then  $\beta/d = 1/8$ , so

$n^{-\left(\frac{\beta_g}{2\beta_g+d}+\frac{\beta_b}{d+2\beta_b}+\frac{\beta_p}{d+2\beta_p}\right)} = n^{-\frac{4\beta}{4\beta+d}} = n^{-1/3}$ . However,  $E_\theta \left[ \tau_{2,k}(\hat{\theta}) - \tau^\dagger \right]$  converges to zero at rate  $n^{-\frac{\beta_b}{d+2\beta_b}} = n^{-\frac{1}{6}}$ . Next

$$\begin{aligned} & var_{\hat{\theta}(\tau^\dagger)} \left\{ \mathbb{ES}_{2,\tilde{\tau}_k(\cdot)}^{test} \left( \hat{\theta}(\tau^\dagger) \right) \right\}^{-1} E_\theta \left[ \mathbb{ES}_{2,\tilde{\tau}_k}^{test} \left( \hat{\theta}(\tau^\dagger) \right) \right] \\ &= O_p \left( n^{-\left(\frac{\beta_b}{d+2\beta_b}+\frac{2\beta_p}{d+2\beta_p}\right)} \right) = n^{-1/6} \gg O_p \left( n^{-\frac{4\beta}{4\beta+d}+\sigma} \right) = n^{-1/3+\sigma} \end{aligned}$$

for small  $\sigma$ . We conclude that the intervals based on  $\mathbb{ES}_{2,\tilde{\tau}_k(\cdot)}^{test} \left( \hat{\theta}(\tau) \right)$  and  $\tau_{2,k}(\hat{\theta}) - \tau$  fail to cover  $\tilde{\tau}_k(\theta) = \tau^\dagger$  at the nominal  $1 - \alpha$  level uniformly over  $\Theta$  as  $n \rightarrow \infty$ . We reach the identical conclusion with regard to the parameter  $\tau(\theta)$  because under our assumptions  $|\tau(\theta) - \tilde{\tau}_k(\theta)| = O_p \left( n^{-\frac{4\beta}{4\beta+d}+\sigma} \right)$

Furthermore, by the argument used in the proof of theorem 42, it is easy to see that the length of each interval is  $O_p(k/n^2) = O_p \left( n^{-\frac{4\beta}{4\beta+d}+\sigma} \right)$ . It follows that if we try to improve the coverage of the intervals based on  $\mathbb{ES}_{2,\tilde{\tau}_k(\cdot)}^{test} \left( \hat{\theta}(\tau) \right)$  and  $\tau_{2,k}(\hat{\theta}) - \tau$  by further increasing  $k$ , the length of the intervals will increase beyond  $O_p \left( n^{-\frac{4\beta}{4\beta+d}+\sigma} \right)$ . We conclude that the interval based on  $\psi_{2,k}(\tau, \hat{\theta})$  is strictly preferred to the other two intervals when  $\beta_p/d = 0$ ,  $\beta_b/d = \beta_g/d = 1/4$  and is never worse in terms of shrinkage rate and coverage than the other two intervals whatever be  $\beta_p$ ,  $\beta_b$ , and  $\beta_g$ . We reach the identical conclusion with regard to the coverage of the parameter  $\tau(\theta)$  because, under our assumptions including our choice of  $k$ ,  $|\tau(\theta) - \tilde{\tau}_k(\theta)| = O_p \left( n^{-\frac{4\beta}{4\beta+d}} \right)$  and  $n^{-\frac{4\beta}{4\beta+d}} \ll n^{-\frac{4\beta}{4\beta+d}+\sigma}$ , the order of the interval lengths.

These results translate directly into analogous results concerning the associated

estimators. Under our assumptions the estimator solving  $\psi_{2,k}(\tau, \hat{\theta}) = 0$  converges to both  $\tau(\theta)$  and  $\tilde{\tau}_k(\theta)$  at rate  $O_p\left(n^{-\frac{4\beta}{4\beta+d}+\sigma}\right)$ . In contrast the rate of convergence of  $\tau_{2,k}(\hat{\theta})$  and the estimator solving  $\mathbb{ES}_{2,\tilde{\tau}_k(\cdot)}^{test}(\hat{\theta}(\tau)) = 0$  converge to  $\tau(\theta)$  and  $\tilde{\tau}_k(\theta)$  at the rates given in (iv) and (v) of theorem 44.

What is the intuition behind the above findings? First note that, as promised by Theorem 2 and part (vii) of the theorem in the last subsection, the bias away from zero of  $var_{\hat{\theta}(\tau^\dagger)}\left\{\mathbb{ES}_{2,\tilde{\tau}_k(\cdot)}^{test}(\hat{\theta}(\tau^\dagger))\right\}^{-1}E_\theta\left[\mathbb{ES}_{2,\tilde{\tau}_k}^{test}(\hat{\theta}(\tau^\dagger))\right]$ ,  $E_\theta\left[\tau_{2,k}(\hat{\theta}) - \tilde{\tau}_k(\theta)\right]$ , and  $v(\hat{\theta})^{-1}E_\theta\left[\psi_{2,k}(\tau^\dagger, \hat{\theta})\right]$  are all  $O_p\left(\left\|\hat{\theta} - \theta\right\|^3\right)$ . However the nature and convergence rate of the  $O_p\left(\left\|\hat{\theta} - \theta\right\|^3\right)$  term can vary markedly between estimators, attaining a minimum for  $E_\theta\left[\psi_{2,k}(\tau^\dagger, \hat{\theta})\right]$ . Now it is not surprising that, for the same order of variance, the order of  $E_\theta\left[\tau_{2,k}(\hat{\theta}) - \tilde{\tau}_k(\theta)\right]$  often exceeds that of  $E_\theta\left[\psi_{2,k}(\tau^\dagger, \hat{\theta})\right]$ . Confidence intervals for  $\tilde{\tau}_k(\theta)$  based on  $\tau_{2,k}(\hat{\theta})$  are centered at (i.e are symmetric around)  $\tau_{2,k}(\hat{\theta})$ , which is a quite stringent constraint on the form of the interval. In that sense, intervals based on  $\tau_{2,k}(\hat{\theta})$  are a higher order generalization of the first order asymptotic Wald intervals for  $\tilde{\tau}_k(\theta)$ . It is well known that when  $\tilde{\tau}_k(\theta)$  is an implicit parameter that sets a functional such as  $\tilde{\psi}_k(\tau, \theta)$  to zero, first-order Wald confidence intervals are often outperformed in finite samples by confidence sets obtained by inverting a 'score-like' test based on first order 'estimating functions' for the functional that depend on the parameter  $\tilde{\tau}_k$  and, frequently, on estimated nuisance parameters as well, although this fact is not reflected in the first order asymptotics. Our example is higher order version



of this phenomenon, where the benefit of the interval  $C_{1-\alpha, \tilde{\psi}_k(\tau)}$  obtained by inverting tests based on the estimating function  $\psi_{2,k}(\tau, \hat{\theta})$  for the functional  $\tilde{\psi}_k(\tau, \theta)$  is clearly and quantitatively revealed by the asymptotics. Note that, like first order Wald intervals, the interval based on  $\tau_{2,k}(\hat{\theta})$  will differ from the interval for  $\tilde{\tau}_k(\theta)$  based on applying an inverse nonlinear monotone transform  $h^{-1}(\cdot)$  to the end points of a Wald interval for the transformed parameter  $h\{\tilde{\tau}_k(\theta)\}$  that is centered on  $h(\tau)_{2,k}(\hat{\theta}) \equiv h(\tilde{\tau}_k(\hat{\theta})) + \mathbb{E}_{2,h(\tilde{\tau}_k(\cdot))}(\hat{\theta})$ . In contrast, like first order score-based intervals, the intervals based on  $\psi_{2,k}(\tau, \hat{\theta})$  and  $\mathbb{E}_{2,\tilde{\tau}_k(\cdot)}^{test}(\hat{\theta}(\tau^\dagger))$  are invariant to monotone transformations of the parameter  $\tilde{\tau}_k(\theta)$ .

More interesting and perhaps more surprising is that, for the same order of variance, the order of  $E_\theta \left[ \text{var}_{\hat{\theta}(\tau^\dagger)} \left\{ \mathbb{E}_{2,\tilde{\tau}_k(\cdot)}^{test}(\hat{\theta}(\tau^\dagger)) \right\}^{-1} \mathbb{E}_{2,\tilde{\tau}_k}^{test}(\hat{\theta}(\tau^\dagger)) \right]$  exceeds that of  $E_\theta \left[ \psi_{2,k}(\tau^\dagger, \hat{\theta}) \right]$ . The surprise derives from a failure to recognize that the theorem 38 is simply too general to help select among competing procedures. For example, this theorem implies that under law  $\hat{\theta}(\tau^\dagger)$ , (a) the variance  $\text{var}_{\hat{\theta}(\tau^\dagger)} \left\{ \mathbb{E}_{2,\tilde{\tau}_k(\cdot)}^{test}(\hat{\theta}(\tau^\dagger)) \right\}^{-1}$  of  $\left[ \text{var}_{\hat{\theta}(\tau^\dagger)} \left\{ \mathbb{E}_{2,\tilde{\tau}_k(\cdot)}^{test}(\hat{\theta}(\tau^\dagger)) \right\}^{-1} E_\theta \left[ \mathbb{E}_{2,\tilde{\tau}_k}^{test}(\hat{\theta}(\tau^\dagger)) \right] \right]$  is less (and generally strictly less) than the variance of  $v(\hat{\theta}(\tau^\dagger))^{-1} E_\theta \left[ \psi_{2,k}(\tau^\dagger, \hat{\theta}(\tau^\dagger)) \right]$ , while (b) both have bias of  $O_p \left( \left\| \hat{\theta}(\tau^\dagger) - \theta \right\|^3 \right)$ . At first blush, this might suggest that the estimator solving  $\mathbb{E}_{2,\tilde{\tau}_k}^{test}(\hat{\theta}(\tau)) = 0$  would likely have the same bias but smaller variance than the estimator solving  $\psi_{2,k}(\tau, \hat{\theta}) = 0$ . But we have seen that just the opposite is true. The reason is that the difference between the variances in (a) is negligible in the sense

that their ratio is  $1 + o_p(1)$ , while the  $O_p\left(\left\|\widehat{\theta}(\tau^\dagger) - \theta\right\|^3\right)$  biases are often of quite different orders with that of  $v\left(\widehat{\theta}(\tau^\dagger)\right)^{-1} E_\theta\left[\psi_{2,k}\left(\tau^\dagger, \widehat{\theta}(\tau^\dagger)\right)\right]$  always a minimum.

More generally, whenever the functional  $\psi(\tau, \theta)$  is in our doubly robust class, Eq 37 holds so  $\widehat{\psi}_{\mathcal{K}_J}^{eff}$  is rate minimax (or near minimax if  $\sigma$  is chosen positive), and the suppositions of Theorem 36 hold for  $\widetilde{\psi}(\tau) = \widehat{\psi}_{\mathcal{K}_J}^{eff}(\tau)$ , Theorem 36 then implies the width of the interval estimator for  $\tau(\theta)$  based on  $\widehat{\psi}_{\mathcal{K}_J}^{eff}(\tau)$  converges to zero at the convergence rate of  $\widehat{\psi}_{\mathcal{K}_J}^{eff}(\tau)$  to  $\psi(\tau, \theta)$ .

## 7 Monotone Missing Data and Other Complex Functionals

### 7.1 Derivation of Higher Order Influence Functions for product of functionals

In this section, we discuss the construction of higher order influence functions for a more general class of functionals than the one thus far considered. An important application of this construction is in the derivation of higher order influence functions in monotone missing data problems. To begin, we must learn to construct higher order influence functions for a functional with the product form:

$$\psi(\theta; \zeta) = \prod_{s=1}^{\zeta} \psi_s(\theta)$$

here  $\psi_s(\theta)$ ,  $s = 1, \dots, \zeta$ , are known to be higher order pathwise differentiable functionals and  $\zeta$  is a known constant.

The following lemma gives the general form of higher order influence functions of  $\psi(\theta; \zeta)$  as a function of influence functions of  $\{\psi_s(\theta) : s = 1, \dots, \zeta\}$ . Before stating our lemma, we need additional notation. Given an ordered set of  $m$  positive integers  $\bar{\mathbf{i}}_m = \{i_1, i_2, \dots, i_m\}$ , for any  $r$  non-negative integers  $\{t_1, t_2, \dots, t_r\}$ , satisfying  $\sum_{s \leq r} t_s = m$ , we define  $\Upsilon = (\bar{i}_{1,t_1}, \bar{i}_{2,t_2}, \dots, \bar{i}_{r,t_r})$  to be an ordered partition of degree  $m$  and order  $r$  if and only if for  $1 \leq s^* \leq r$  and  $t_{s^*} > 0$  we have:

$$\bar{i}_{s^*, t_{s^*}} = \left\{ \begin{array}{l} i_{j(s^*)+1}, i_{j(s^*)+2}, \dots, i_{j(s^*)+t_{s^*}} : \\ \\ j(s^*) = \\ \left\{ \begin{array}{ll} \sum_{q=1}^{s^*-1} t_q & \text{for } 1 < s^* \leq r \\ 0 & \text{for } s^* = 1 \end{array} \right. \end{array} \right\}$$

and for all  $t_{s^*} = 0$  we have  $\bar{i}_{s^*, t_{s^*}} = \emptyset$ . Any such ordered partition satisfies

$$\bar{\mathbf{i}}_m = \bigcup_{s=1}^r \bar{i}_{s, t_s}$$

**Lemma 45** *Let  $if_{\psi_s(\theta); j, j}(o_{i_1}, \dots, o_{i_j}; \theta) = IF_{\psi_s(\theta); j, j; \bar{\mathbf{i}}_j}(\theta)$  for  $j \geq 1$  be the  $j$ th order influence function of  $\psi_s(\theta)$ , and define  $IF_{\psi_s(\theta); 0, 0; \bar{\mathbf{i}}_0}(\theta) \equiv \psi_s(\theta)$ . Then the  $m$ th order influence function of  $\psi(\theta; \zeta)$  is given by:*

$$\begin{aligned} \mathbb{IF}_{\psi(\theta; \zeta), m}(\theta) &= \sum_{j=1}^m \mathbb{IF}_{\psi(\theta; \zeta), j, j}(\theta) \\ &= \sum_{j=1}^m \mathbb{V} \left[ IF_{\psi(\theta; \zeta), j, j; \bar{\mathbf{i}}_j}(\theta) \right] \end{aligned}$$

where

$$\mathbb{V} \left[ IF_{\psi(\theta; \zeta); j, j, \bar{\mathbf{i}}_j}(\theta) \right] = \mathbb{V} \left[ \sum_{\{t_1, \dots, t_\zeta\} \in \Upsilon_{\zeta; j}} \prod_{s=1}^{\zeta} IF_{\psi_s(\theta); t_s, t_s, \bar{\mathbf{i}}_{s, t_s}}(\theta) \right]$$

$$\Upsilon_{\zeta; j} = \left\{ t_1, \dots, t_\zeta : \sum_{s=1}^{\zeta} t_s = j, t_s \geq 0 \right\}$$

It is easy to generalize this lemma to functionals of the more general form

$$\psi(\theta) = \psi(\theta; \zeta_1, \zeta_2) = \sum_{1 \leq v_2 \leq \zeta_1} \prod_{s=1}^{\zeta_2} \psi_{v_2, s}(\theta)$$

where both  $\zeta_1, \zeta_2$  are known constants, since the higher order influence function of a linear combination of functionals is the linear combination of the influence functions of the functionals.

## 7.2 Application to Monotone Missing Data

### 7.2.1 Mapping Higher Order Full Data IFs to Observed Data IFs

We next turn to the analysis of missing data models. Suppose that we have derived the  $m$ th order influence function  $\mathbb{IF}_{\tilde{\psi}_k, m}^{full}(\theta)$   $m \geq 2$  for a truncated parameter  $\tilde{\psi}_k$  of a parameter  $\psi(\theta)$  in our doubly robust class of functionals based on i.i.d full data  $L_{full}$ ; then, according to theorem 45 the estimated  $\mathbb{IF}_{\tilde{\psi}_k; j, j}^{full}(\hat{\theta})$ ,  $j \leq m$ , is of the form

$$\begin{aligned} \sum_{1 \leq l \leq k^{(j-1)}} \mathbb{V} \left[ \prod_{s=1}^j \hat{U}_{l, s, i_s}^{(j)} \right] &= \sum_{1 \leq l \leq k^{(j-1)}} \left\{ \mathbb{V} \left( \prod_{s=1}^j u_{l, s}^{(j)}(L_{i_s}^{full}; \hat{\theta}) \right) \right\} \\ &= \sum_{1 \leq l \leq k^{(j-1)}} \mathbb{V} \left[ \prod_{s=1}^j \hat{U}_{l, s, i_s}^{(j)} \right] \end{aligned}$$

where  $u_{l,s}^{(j)}(L_{i_s}^{full}; \hat{\theta}) \equiv \hat{U}_{l,s,i_s}^{(j)}$  depends of one subject's data and has a functional form which depends on the form of  $H(\cdot, \cdot)$  corresponding to the functional of interest. For example,  $\hat{U}_{1,1,i_1}^{(2)} = \left[ (A - \hat{P}) Z_1 \right]_{i_1}$  corresponds to the first component of the vector contributed by person  $i_1$  to  $\mathbb{IF}_{\psi_k;2,2}^{full}(\hat{\theta})$ . Note that  $E_{\hat{\theta}}[\hat{U}_{l,s}^{(j)}] = 0$ . Next, suppose that for a subject some of the data may be missing, so that we observe

$$\{(L_{obs,i} = R_i L_{full,i} + (1 - R_i)w(L_{full,i}), R_i) : i = 1, \dots, n\}$$

where  $R = I[L_{obs,i} = L_{full,i}]$  instead of the full data  $L_{full,i}$ . For now,  $W_i = w(L_{full,i})$  is a known function of the full data and  $W_i$  is always observed. Below we shall extend our results to monotone missing data.

Define

$$\pi(W; \theta) = P(R = 1 | L^{full}; \theta)$$

$$B_{l,s}^{(j)}(W) = E(\hat{U}_{l,s}^{(j)} | W; \theta).$$

and we suppose  $\pi(W; \theta) > \sigma > 0$ . If we know  $\pi(W; \theta)$ , we can use

$$\mathbb{V} \left[ \prod_{s=1}^j \left( \frac{R_{i_s} \hat{U}_{l,s,i_s}^{(j)}}{\pi(L_{obs,,i_s}; \theta)} + \phi(L_{obs,,i_s}; \theta) \right) \right]$$

instead of  $\mathbb{V} \left[ \prod_{s=1}^j \hat{U}_{l,s,i_s}^{(j)} \right]$ , where  $\phi(\cdot)$  is an arbitrary function with finite variance which satisfies  $E_{\theta}[\phi(L_{obs}) | L_{full}] = 0$ . It is easy to verify that this statistic is a function of the observed data and an unbiased estimator of

$$E_{\theta} \left[ \mathbb{V} \left[ \prod_{s=1}^j \hat{U}_{l,s,i_s}^{(j)} \right] \right]$$

so that in fact, in terms of rates, our observed data statistic has the same bias and variance properties as the original full data higher order influence function. Moreover, the most efficient (in terms of constants) choice for  $\phi(L_{obs}; \theta)$  in our class of estimators is  $-\left(\frac{R}{\pi(L_{obs})} - 1\right) B_{l,s}^{(j)}(L_{obs}; \theta)$ , while  $B_{l,s}^{(j)}(L_{obs}; \theta)$  is an unknown function to be estimated from the observed data. This last two observations motivate our proposed strategy for mapping higher order influence functions in the full data model to higher order influence functions in the observed data model when the missingness mechanism is unknown. First, define

$$\begin{aligned} \prod_{s=1}^j \tau_{l,s}^{(j)}(\theta) &\equiv E_{\theta} \left[ \mathbb{V} \left[ \prod_{s=1}^j \widehat{U}_{l,s,i_s}^{(j)} \right] \right] \\ &= E_{\theta} \left[ \mathbb{V} \left[ \prod_{s=1}^j \left( \frac{R_{i_s} [\widehat{U}_{l,s,i_s}^{(j)} - B_{l,s}^{(j)}(L_{obs,i_s}; \theta)]}{\pi(L_{obs,i_s}; \theta)} + B_{l,s}^{(j)}(L_{obs,i_s}; \theta) \right) \right] \right] \end{aligned}$$

which we notice is of the product form that was discussed earlier in this section. In order to construct an  $m$ th order influence function for  $\tilde{\psi}_k$ , it is now apparent that we must first succeed in constructing  $m$ th order influence functions for the set of parameters  $\left\{ \prod_{s=1}^j \tau_{l,s}^{(j)}(\theta) : j \leq m, l \leq k^{(j-1)} \right\}$  so as to guarantee an estimation bias of order  $m+1$ . Now, for each  $l$  and  $s$ ,  $\tau_{l,s}^{(j)}(\theta)$  is itself a member of our general doubly robust class so that  $\tau_{l,s}^{(j)}(\theta) = E_{\theta} [H^{(j)}(P_{l,s}, B_{l,s}^{(j)})]$  where  $H_{l,s,1} = -R, H_{l,s,2} = 1, H_{l,s,3} = R\widehat{U}_{l,s}, H_{l,s,4} = 0$ , and  $P_{l,s} = \pi(L_{obs,i_s}; \theta)^{-1}, B_{l,s}^{(j)} = B_{l,s}^{(j)}(L_{obs,i_s}; \theta)$ . Thus, it immediately follows that neither  $\left\{ \prod_{s=1}^j \tau_{l,s}^{(j)}(\theta) : j \leq m, l \leq k^{(j-1)} \right\}$  nor  $\tilde{\psi}_k$  have higher order influence functions. We proceed by constructing influence functions for the set

of truncated parameters  $\left\{ \prod_{s=1}^j \tilde{\tau}_{l,s}^{(j)}(\theta) : j \leq m, l \leq k^{(j-1)} \right\}$  where  $\tilde{\tau}_{l,s}^{(j)}(\theta)$  are appropriately truncated versions of  $\tau_{l,s}^{(j)}(\theta)$  as in section 3.2.2. We then define the truncated parameter  $\tilde{\psi}_{k,m}$

$$\tilde{\psi}_{k,m}(\theta) = \psi(\hat{\theta}) + \sum_{1 \leq v \leq m} \sum_{1 \leq l \leq k^{(v-1)}} \prod_{s=1}^v \tilde{\tau}_{l,s}^{(v)}(\theta)$$

Note that  $\tilde{\psi}_{k,m}(\theta)$  differs from the truncated parameters  $\tilde{\psi}_k(\theta)$  of section 3.2.2, in that it depends on the order of the influence function we plan to base our inference on. The next theorem gives the  $m$ th order influence function of  $\tilde{\psi}_{k,m}(\theta)$ .

**Theorem 46** *Let  $IF_{\tilde{\tau}_{l,s}^{(j)}(\theta);0,0;\bar{i}_{s,0}}(\theta) \equiv \tilde{\tau}_{l,s}^{(j)}(\theta)$ , The  $m$ th order influence function of  $\tilde{\psi}_{k,m}(\theta)$  is given by*

$$\mathbb{IF}_{\tilde{\psi}_{k,m}(\theta),m}(\theta) = \sum_{j=1}^m \mathbb{V} \left[ IF_{\tilde{\psi}_{k,m}(\theta),j,j;\bar{i}_j}(\theta) \right]$$

where

$$IF_{\tilde{\psi}_{k,m}(\theta),j,j;\bar{i}_j}(\theta) = \sum_{1 \leq v \leq j} \sum_{1 \leq l \leq k^{(v-1)}} \sum_{\{t_1, \dots, t_v\} \in \Upsilon_{v,j}} \prod_{s=1}^v IF_{\tilde{\tau}_{l,s}(\theta);t_s,t_s;\bar{i}_{s,t_s}}(\theta) \quad (54)$$

$$\Upsilon_{v,j} = \left\{ t_1, \dots, t_v : \sum_{s=1}^v t_s = j, t_s \geq 0 \right\}$$

**Proof.** The proof follows directly from the lemma 45 applied to each element

$$\left\{ \prod_{s=1}^v \tilde{\tau}_{l,s}^{(v)}(\theta) : 1 \leq v \leq j, 1 \leq l \leq k^{(v-1)}, 1 \leq s \leq v \right\}$$

■

**Corollary 47** *The  $m$ th order estimated influence function of  $\tilde{\psi}_{k,m}(\theta)$  is given by*

$$\mathbb{IF}_{\tilde{\psi}_{k,m}(\theta),m}(\hat{\theta}) = \sum_{j=1}^m \mathbb{V} \left[ IF_{\tilde{\psi}_{k,m}(\theta),j,j;\bar{\mathbf{i}}_j}(\hat{\theta}) \right]$$

where

$$IF_{\tilde{\psi}_{k,m}(\theta),j,j;\bar{\mathbf{i}}_j}(\hat{\theta}) = \left[ \sum_{1 \leq v \leq m} \sum_{1 \leq l \leq k(v-1)} \sum_{\{t_1, \dots, t_v\} \in \Upsilon_{v;j}^+} \prod_{s=1}^v IF_{\tilde{\tau}_{l,s}(\theta);t_s,t_s;\bar{\mathbf{i}}_{s,t_s}}(\hat{\theta}) \right]$$

$$\Upsilon_{v;j}^+ = \left\{ t_1, \dots, t_v : \sum_{s=1}^v t_s = j, t_s > 0 \right\}$$

**Proof.** This result follows immediately from theorem 46 and the fact that

$$IF_{\tilde{\tau}_{l,s}(\theta);0,0;\bar{\mathbf{i}}_{s,0}}(\hat{\theta}) \equiv \tilde{\tau}_{l,s}(\hat{\theta}) = 0$$

by definition. ■

### 7.2.2 Two-occasion Monotone Missing Data

We are now ready to use our theorem to derive higher order influence functions for a functional  $\psi(\theta)$  from monotone missing data. We begin with the simple two-occasion case. Let  $D_i = (L_{0,i}, L_{1,i}, Y_i) \sim F(D_i, \vartheta)$ ,  $i = 1, \dots, n$ , be  $n$  *i.i.d* copies constitute the full data. The outcome of interest,  $Y_i$ , is univariate.  $L_{0,i}$  and  $L_{1,i}$  are vectors of continuous covariates with dimensions  $d_0$  and  $(d_1 - d_0)$  respectively. The observed data  $O_i = (R_{0,i}, L_{0,i}, R_{0,i}L_{1,i}, R_{1,i}, R_{0,i}R_{1,i}Y_i) \sim F(O_i, \theta = (\vartheta, \gamma))$ , where both  $R_{0,i}$  and  $R_{1,i}$  are binary missing indicators. The outcome  $Y$  is observed if and only if  $R_0$



$= R_1 = 1$ , while  $L_1$  is observed if and only if  $R_0 = 1$ . The data is assumed to be missing at random, that is:

$$\Pr(R_0 = 1|D) = \Pr(R_0 = 1|L_0) = \pi_0(L_0; \theta)$$

and

$$\begin{aligned} \Pr(R_1 = 1|R_0 = 1, D) &= \Pr(R_1 = 1|R_0 = 1, L_0, L_1) \\ &= \pi_1(L_0, L_1; \theta) \end{aligned}$$

Under this monotone missing-data pattern, the parameter of interest is given by

$$\psi(\theta) = E_\theta(Y) = E_\theta\left(\frac{R_0 R_1}{\pi_0(L_0; \theta) \pi_1(L_0, L_1; \theta)} Y\right)$$

We impose the following assumptions: (a1).  $B_0 = b_0(L_0; \theta) = E_\theta[Y|L_0] \in H(\beta_{b_0}, C_{B_0})$ ; (a2).  $B_1 = b_1(L_0, L_1; \theta) = E_\theta[Y|L_0, L_1] \in H(\beta_{b_1}, C_{B_1})$ ; (b1).  $\pi_0(L_0; \theta) = \Pr(R_0 = 1|L_0) \in H(\beta_{\pi_0}, C_{\pi_0})$  and  $0 < \sigma_{\pi_0}^l < \pi_0(L_0; \theta)$  w.p. 1 (b2).  $\pi_1(L_0, L_1; \theta) = \Pr(R_1 = 1|L_0, L_1) \in H(\beta_{\pi_1}, C_{\pi_1})$  and  $0 < \sigma_{\pi_1}^l < \pi_1(L_0, L_1; \theta)$  w.p. 1. (c1). The marginal density of  $L_0$ ,  $f_0 = f(L_0; \theta)$ , falls in a Hölder ball  $H(\beta_{f_0}, C_{f_0})$ , and  $0 < \sigma_{f_0} < f_0(L_0; \theta)$  w.p. 1,  $\|f_0\|_\infty \leq C_{f_0}^* < \infty$ . (c2). The marginal density of  $(L_0, L_1)$ ,  $f_1 = f(L_0, L_1; \theta)$ , falls in a Hölder ball  $H(\beta_{f_1}, C_{f_1})$ , and  $0 < \sigma_{f_1} < f_1(L_0, L_1; \theta)$  w.p. 1,  $\|f_1\|_\infty \leq C_{f_1}^* < \infty$ .

We define  $g_0(L_0; \theta) = \pi_0 f_0$ ,  $g_1(L_0, L_1; \theta) = \pi_0 \pi_1 f_1$  with corresponding Hölder exponents  $\beta_{g_0} = \min(\beta_{f_0}, \beta_{\pi_0})$  and  $\beta_{g_1} = \min(\beta_{f_1}, \beta_{\pi_0}, \beta_{\pi_1})$  respectively.

We next show how to apply theorem 46 in a nested fashion in order to derive higher order U-statistic estimators  $\{\widehat{\psi}_m : m \geq 1\}$  in the observed data model. In the first step of our procedure, we derive higher order influence functions for a truncated parameter in the artificial missing data problem in which the observed data is given by  $O_i^\dagger = (R_{0,i}, L_{0,i}, R_{0,i}L_{1,i}, R_{0,i}Y_i)$  rather than  $O_i$ . In the second step, we construct influence functions from a second artificial missing data problem with  $O_i^\dagger$  now the full data and  $O_i$  the observed data. This final influence function is, in fact, the influence function for a truncated parameter in the original monotone missing data model.

To follow this nested procedure, we derive a first stage class of estimators

$$\left\{ \widehat{\psi}_m^\dagger : m \geq 1 \right\},$$

which are functions of  $O_i^\dagger$ . In this model,  $D_i$  is the full data,  $O_i^\dagger$  is the observed data,  $R_{0,i}$  is the missing indicator,  $L_{0,i}$  is a vector of always observed covariates, and  $(Y_i, L_{1i})$  is the outcome which might be missing. Since the parameter of interest  $\psi(\theta)$  is the marginal mean of  $Y$ , this is Example 2a of section 3.1. Therefore results from this example may be applied, hence

$$if_{1,\psi(\theta)}^F(D_i) = Y_i - \psi(\theta)$$

and

$$if_{jj,\psi(\theta)}^F(D_i) = 0 \quad \text{for } \forall j \geq 2.$$

Moreover,

$$if_{1,\psi(\theta)}(O_i^\dagger) = \frac{R_{0,i}}{\pi_{0,i}}(Y_i - B_{0,i}) + B_{0,i} - \psi(\theta)$$

so that we can define

$$\begin{aligned}\tilde{B}_0 &= \hat{B}_0 + \bar{Z}_{k_0}^T E_\theta \left[ \frac{R_0}{\hat{\pi}_0} \bar{Z}_{k_0} \bar{Z}_{k_0}^T \right]^{-1} E_\theta \left[ \frac{R_0}{\hat{\pi}_0} (Y - \hat{B}_0) \bar{Z}_{k_0} \right] \\ \tilde{\pi}_0^{-1} &= \hat{\pi}_0^{-1} \left( 1 - \bar{Z}_{k_0}^T E_\theta \left[ \frac{R_0}{\hat{\pi}_0} \bar{Z}_{k_0} \bar{Z}_{k_0}^T \right]^{-1} E_\theta \left[ \left( \frac{R_0}{\hat{\pi}_0} - 1 \right) \bar{Z}_{k_0} \right] \right) \\ \tilde{\psi}^\dagger(\theta) &= E_\theta \left[ \frac{R_0}{\tilde{\pi}_0} (Y - \tilde{B}_0) + \tilde{B}_0 \right]\end{aligned}$$

with  $\dot{B} = 1$  and  $\dot{P} = \hat{\pi}_0^{-1}$ .

Moreover

$$\hat{\psi}_m^\dagger = \psi(\hat{\theta}) + \mathbb{V}_n(\widehat{IF}_{m,\tilde{\psi}^\dagger})$$

$(\hat{B}_0, \hat{\pi}_0, \hat{f}(L_0))$  are rate optimal nonparametric estimators of  $(B_0, \pi_0, f_0)$  respectively estimated from the training sample and  $\bar{Z}_{k_0} = \bar{z}_{k_0}(L_0) = \hat{E}(\bar{\varphi}_{k_0}(L_0) \bar{\varphi}_{k_0}^T(L_0))^{-1/2} \bar{\varphi}_{k_0}(L_0)$ .

From theorem 18 and eq.(34) we also know that:

$$\begin{aligned}E(\hat{\psi}_m^\dagger) - \psi(\theta) &= TB_{k_0} + EB_m \\ &= O_P \left( \max \left[ k_0^{-\frac{\beta_{b_0} + \beta_{\pi_0}}{d_0}}, \left( \frac{\log n}{n} \right)^{\frac{(m-1)\beta_{g_0}}{d_0 + 2\beta_{g_0}}} n^{-\left( \frac{\beta_{b_0}}{d_0 + 2\beta_{b_0}} + \frac{\beta_{\pi_0}}{d_0 + 2\beta_{\pi_0}} \right)} \right] \right)\end{aligned}$$

Next, we proceed with the second step of our procedure, where  $O_i^\dagger$  is now the full data and  $O_i$  becomes the observed data. Then,  $O_i^\dagger = O_i$  if  $R_{0,i} = 0$  or  $R_{0,i} = R_{1,i} = 1$ .

Therefore we may define a new missing indicator

$$R = (1 - R_0) + R_0 R_1$$

with

$$\Pr(R_i = 1|O_i) = (1 - R_{0,i}) + R_{0,i}\pi_{1,i}$$

In contrast with the first phase of our procedure, the full data influence function now has non-zero higher order contributions; thus we can proceed with the strategy layed

out in the first part of this section. We can put  $\widehat{IF}_{jj,\tilde{\psi}^\dagger(\theta),\tilde{i}_j}$  in the format of eq.(54)

as

$$\begin{aligned} \widehat{IF}_{jj,\tilde{\psi}^\dagger(\theta),\tilde{i}_j} &= (-1)^{j-1} \sum_{s_1=1}^{k_0} \dots \sum_{s_{j-1}=1}^{k_0} \left\{ \left( \frac{R_0}{\widehat{\pi}_0} (Y - \widehat{B}_0) Z_{s_1} \right)_{i_1} \left( \left( \frac{R_0}{\widehat{\pi}_0} - 1 \right) Z_{s_{j-1}} \right)_{i_j} \times \right. \\ &\quad \left. \prod_{t=2}^{j-1} \left( \frac{R_0}{\widehat{\pi}_0} Z_{s_{t-1}} Z_{s_t} - I(s_{t-1} = s_t) \right)_{i_t} \right\} \\ &= (-1)^{j-1} \sum_{l=1}^{k_0^{j-1}} \left\{ \left( \frac{R_0}{\widehat{\pi}_0} (Y - \widehat{B}_0) Z_{n(l,1)} \right)_{i_1} \left( \left( \frac{R_0}{\widehat{\pi}_0} - 1 \right) Z_{n(l,j-1)} \right)_{i_j} \times \right. \\ &\quad \left. \prod_{t=2}^{j-1} \left( \frac{R_0}{\widehat{\pi}_0} Z_{n(l,t-1)} Z_{n(l,t)} - I(n(l,t-1) = n(l,t)) \right)_{i_t} \right\} \end{aligned}$$

where  $n(l) : \{1, 2, \dots, k_0^{j-1}\} \rightarrow \{1, 2, \dots, k_0\}^{j-1}$  is a one-to-one mapping that indexes all permutations of  $\{(s_1, s_2, \dots, s_{j-1}), 1 \leq s_t \leq k_0\}$ , and  $n(l, t)$  is the  $t$ th entry of  $n(l)$ .

We define  $\forall j > 1$ .

$$\widehat{U}_{l,s}^{(j)} \equiv \begin{cases} \frac{R_0}{\widehat{\pi}_0} (Y - \widehat{B}_0) Z_{n(l,1)} & \text{for } s = 1 \\ \left\{ \begin{array}{c} \frac{R_0}{\widehat{\pi}_0} Z_{n(l,s-1)} Z_{n(l,s)} \\ -I(n(l,s-1) = n(l,s)) \end{array} \right\} & \text{for } 1 < s < j \\ (-1)^{j-1} \left( \frac{R_0}{\widehat{\pi}_0} - 1 \right) Z_{n(l,j-1)} & \text{for } s = j \end{cases}$$

$$\tau_{l,s}^{(j)}(\theta) \equiv E_\theta \left( \widehat{U}_{l,s}^{(j)} \right)$$

Let  $V_i = (R_{0,i}, L_{0,i}, R_{0,i}L_{1,i})$  which is always observed. Then

$$\begin{aligned}\tau_{l,1}^{(j)}(\theta) &= E_\theta \left( \frac{R_0}{\widehat{\pi}_0} (Y - \widehat{B}_0) Z_{n(l,1)} \right) \\ &= E_\theta \left( \frac{R}{\pi} \left[ \begin{array}{c} \frac{R_0}{\widehat{\pi}_0} (Y - \widehat{B}_0) Z_{n(l,1)} - \\ E \left( \frac{R_0}{\widehat{\pi}_0} (Y - \widehat{B}_0) Z_{n(l,1)} | V \right) \end{array} \right] + E_\theta \left( \frac{R_0}{\widehat{\pi}_0} (Y - \widehat{B}_0) Z_{n(l,1)} | V \right) \right) \\ &= E_\theta \left( \frac{R_1}{\pi_1(\theta)} \frac{R_0}{\widehat{\pi}_0} (Y Z_{n(l,1)} - B_1(\theta) Z_{n(l,1)}) + \frac{R_0}{\widehat{\pi}_0} (B_1(\theta) Z_{n(l,1)} - \widehat{B}_0 Z_{n(l,1)}) \right)\end{aligned}$$

Thus,  $\tau_{l,1}^{(j)}(\theta)$  falls into the doubly-robust  $H(b, p)$  class of functionals with  $B^{\tau_{l,1}^{(j)}} = B_1(\theta) Z_{n(l,1)}$ ,  $P^{\tau_{l,1}^{(j)}} = \pi_1^{-1}(\theta)$ , and corresponding  $H_1^{\tau_{l,1}^{(j)}} = -R_1 \frac{R_0}{\widehat{\pi}_0}$ ,  $H_2^{\tau_{l,1}^{(j)}} = \frac{R_0}{\widehat{\pi}_0}$ ,  $H_3^{\tau_{l,1}^{(j)}} = R_1 \frac{R_0}{\widehat{\pi}_0} Y Z_{n(l,1)}$ ,  $H_4^{\tau_{l,1}^{(j)}} = -\frac{R_0}{\widehat{\pi}_0} \widehat{B}_0 Z_{n(l,1)}$ . For any  $s > 1$ ,  $\widehat{U}_{l,s}^{(j)} = \widehat{u}_{l,s}^{(j)}(O)$  is a known function of the observed data, thus  $if_{1,\tau_{l,1}^{(j)}(\theta)} = \widehat{u}_{l,s}^{(j)}(O) - \tau_{l,1}^{(j)}(\theta)$  and  $if_{m,m,\tau_{l,1}^{(j)}(\theta)} = 0$  for any  $m > 1$ . Moreover, choosing  $\dot{B}^{\tau_{l,1}^{(j)}} = 1$ ,  $\dot{P}^{\tau_{l,1}^{(j)}} = \widehat{\pi}_1^{-1}$ , so that:

$$\begin{aligned}\widetilde{B_1 Z_{n(l,1)}} &= \left\{ \begin{array}{c} \widehat{B}_1 Z_{n(l,1)} + \overline{W}_{k_1}^T E_\theta \left( \frac{R_1}{\widehat{\pi}_1} \frac{R_0}{\widehat{\pi}_0} \overline{W}_{k_1} \overline{W}_{k_1}^T \right)^{-1} \\ \times E_\theta \left( \frac{R_1}{\widehat{\pi}_1} \frac{R_0}{\widehat{\pi}_0} (Y - \widehat{B}_1) Z_{n(l,1)} \overline{W}_{k_1} \right) \end{array} \right\} \\ \widetilde{\pi_1^{-1}} &= \left\{ \widehat{\pi}_1^{-1} \left( \begin{array}{c} 1 - \overline{W}_{k_1}^T E_\theta \left[ \frac{R_1}{\widehat{\pi}_1} \frac{R_0}{\widehat{\pi}_0} \overline{W}_{k_1} \overline{W}_{k_1}^T \right]^{-1} \\ \times E_\theta \left[ \frac{R_0}{\widehat{\pi}_0} \left( \frac{R_1}{\widehat{\pi}_1} - 1 \right) \overline{W}_{k_1} \right] \end{array} \right) \right\}\end{aligned}$$

where  $\overline{W}_{k_1} = \overline{w}_{k_1}(L_0, L_1) = \widehat{E}(\overline{\varphi}_{k_1}(L_0, L_1) \overline{\varphi}_{k_1}^T(L_0, L_1))^{-1/2} \overline{\varphi}_{k_1}(L_0, L_1)$ , and  $\overline{\varphi}_{k_1}(L_0, L_1)$

is a  $k_1$ -dimensional vector of tensor product basis for functions of  $(L_0, L_1)$ . From the-

orem 46, for  $1 \leq l \leq k_0^{j-1}$ :

$$\begin{aligned} & \tilde{\tau}_{l,1}^{(j)}(\theta) \\ & \equiv E_{\theta} \left( \begin{aligned} & R_1 \tilde{\pi}_1^{-1}(\theta) \frac{R_0}{\hat{\pi}_0} \left( Y Z_{n(l,1)} - B_1(\theta) \widetilde{Z_{n(l,1)}} \right) \\ & + \frac{R_0}{\hat{\pi}_0} \left( B_1(\theta) \widetilde{Z_{n(l,1)}} - \hat{B}_0 Z_{n(l,1)} \right) \end{aligned} \right) \end{aligned}$$

have higher order influence functions  $\left\{ \mathbb{IF}_{m, \tilde{\tau}_{l,1}^{(j)}}(\theta), m \geq 1 \right\}$ . Let  $\tilde{\tau}_{l,t}^{(j)}(\theta) \equiv \tau_{l,t}^{(j)}(\theta)$  for any  $t > 1$ , and  $\left\{ \mathbb{IF}_{m, \tilde{\psi}_{j,j}}(\theta), m \geq 1 \right\}$  be the higher order influence functions of  $\tilde{\psi}_{j,j}(\theta) = \sum_{l=1}^{k_0^{j-1}} \prod_{t=1}^j \tilde{\tau}_{l,t}^{(j)}(\theta)$ .

For  $j = 1$ , define

$$\begin{aligned} \tau_{1,1}^{(1)} &= E_{\theta} \left( \widehat{IF}_{1, \tilde{\psi}^{\dagger}} \left( O_i^{\dagger} \right) \right) \\ &= E_{\theta} \left( \frac{R}{\pi} \left( \widehat{IF}_{1, \tilde{\psi}^{\dagger}} \left( O_i^{\dagger} \right) - E_{\theta} \left( \widehat{IF}_{1, \tilde{\psi}^{\dagger}} \left( O_i^{\dagger} \right) | V_i \right) \right) + E_{\theta} \left( \widehat{IF}_{1, \tilde{\psi}^{\dagger}} \left( O_i^{\dagger} \right) | V_i \right) \right) \\ &= E_{\theta} \left( \frac{R_1}{\pi_1(\theta)} \frac{R_0}{\hat{\pi}_0} (Y - B_1(\theta)) + \frac{R_0}{\hat{\pi}_0} \left( B_1(\theta) - \hat{B}_0 \right) + \hat{B}_0 - \psi(\hat{\theta}) \right) \end{aligned}$$

$\tau_{1,1}^{(1)}$  also belongs to the  $H(\cdot, \cdot)$  class of models with  $B^{\tau_{1,1}^{(1)}} = B_1$ ,  $P^{\tau_{1,1}^{(1)}} = \pi_1^{-1}$ ,  $H_1 = -R_1 \frac{R_0}{\hat{\pi}_0}$ ,  $H_2 = \frac{R_0}{\hat{\pi}_0}$ ,  $H_3 = R_1 \frac{R_0}{\hat{\pi}_0} Y$ ,  $H_4 = \left( 1 - \frac{R_0}{\hat{\pi}_0} \right) \hat{B}_0 - \psi(\hat{\theta})$ . We may choose  $\dot{B}^{\tau_{1,1}^{(1)}} = B_1$ ,  $\dot{P}^{\tau_{1,1}^{(1)}} = \hat{\pi}_1^{-1}$  so that

$$\begin{aligned} \tilde{B}_1 &= \left\{ \begin{aligned} & \hat{B}_1 + \overline{W}_{k_1}^T E_{\theta} \left( \frac{R_1}{\hat{\pi}_1} \frac{R_0}{\hat{\pi}_0} \overline{W}_{k_1} \overline{W}_{k_1}^T \right)^{-1} \\ & \times E_{\theta} \left( \frac{R_1}{\hat{\pi}_1} \frac{R_0}{\hat{\pi}_0} \left( Y - \hat{B}_1 \right) \overline{W}_{k_1} \right) \end{aligned} \right\} \\ \tilde{\pi}_1^{-1} &= \left\{ \hat{\pi}_1^{-1} \left( \begin{aligned} & 1 - \overline{W}_{k_1}^T E_{\theta} \left[ \frac{R_1}{\hat{\pi}_1} \frac{R_0}{\hat{\pi}_0} \overline{W}_{k_1} \overline{W}_{k_1}^T \right]^{-1} \\ & \times E_{\theta} \left[ \frac{R_0}{\hat{\pi}_0} \left( \frac{R_1}{\hat{\pi}_1} - 1 \right) \overline{W}_{k_1} \right] \end{aligned} \right) \right\} \end{aligned}$$

and

$$\tilde{\tau}_{1,1}^{(1)}(\theta) \equiv E_{\theta} \left( R_1 \tilde{\pi}_1^{-1} \frac{R_0}{\widehat{\pi}_0} (Y - \tilde{B}_1) + \frac{R_0}{\widehat{\pi}_0} (\tilde{B}_1 - \hat{B}_0) + \hat{B}_0 - \psi(\hat{\theta}) \right)$$

has higher order IFs  $\left\{ \mathbb{IF}_{m, \tilde{\tau}_{1,1}^{(1)}(\theta)}, m \geq 1 \right\}$ .

**Theorem 48** Define  $\tilde{\psi}_{k_0, k_1, m}(\theta) = \psi(\hat{\theta}) + \tilde{\tau}_{1,1}^{(1)}(\theta) + \sum_{j=2}^m \tilde{\psi}_{j,j}(\theta)$ , then  $\forall m^* \leq m$

$$\begin{aligned} \mathbb{IF}_{m^*, \tilde{\psi}_{k_0, k_1, m}(\theta)}(\hat{\theta}) &= \mathbb{IF}_{m^*, \tilde{\psi}_{k_0, k_1, m^*}(\theta)}(\hat{\theta}) \\ &= \mathbb{IF}_{1, \tilde{\psi}_{k_0, k_1, m}(\theta)}(\hat{\theta}) + \sum_{j=2}^{m^*} \mathbb{IF}_{j, j, \tilde{\psi}_{k_0, k_1, m}(\theta)}(\hat{\theta}) \end{aligned}$$

with

$$\widehat{IF}_{1, \tilde{\psi}_{k_0, k_1, m}(\theta)}(O_i) = \left\{ \begin{array}{l} \frac{R_{1,i}}{\widehat{\pi}_{1,i}} \frac{R_{0,i}}{\widehat{\pi}_{0,i}} (Y_i - \hat{B}_{1,i}) + \\ \frac{R_{0,i}}{\widehat{\pi}_{0,i}} (\hat{B}_{1,i} - \hat{B}_{0,i}) + \hat{B}_{0,i} - \psi(\hat{\theta}) \end{array} \right\}$$

and  $\forall j \geq 2$

$$\begin{aligned} &\widehat{IF}_{j, j, \tilde{\psi}_{k_0, k_1, m}(\theta)}(O_{i_1}, \dots, O_{i_j}) \\ &= (-1)^{j-1} \sum_{t=2}^{j-1} \left[ \begin{array}{l} \left\{ \begin{array}{l} \left[ \frac{R_0}{\widehat{\pi}_0} \frac{R_1}{\widehat{\pi}_1} (Y - \hat{B}_1) \right]_{i_1} \overline{W}_{k_1, i_1}^T \times \\ \prod_{s=2}^{j-1} \left( \frac{R_0}{\widehat{\pi}_0} \frac{R_1}{\widehat{\pi}_1} \overline{W}_{k_1} \overline{W}_{k_1}^T - I \right)_{i_s} \overline{W}_{k_1, i_j} \left[ \frac{R_0}{\widehat{\pi}_0} \left( \frac{R_1}{\widehat{\pi}_1} - 1 \right) \right]_{i_j} \end{array} \right\} \\ \left\{ \begin{array}{l} \left( \frac{R_0}{\widehat{\pi}_0} \left( \frac{R_1}{\widehat{\pi}_1} - 1 \right) \right)_{i_j} \overline{W}_{k_1, i_j}^T \prod_{l=t+1}^{j-1} \left( \frac{R_1}{\widehat{\pi}_1} \frac{R_0}{\widehat{\pi}_0} \overline{W}_{k_1} \overline{W}_{k_1}^T - I \right)_{i_l} \\ \times \left[ \overline{W}_{k_1} \frac{R_1}{\widehat{\pi}_1} \frac{R_0}{\widehat{\pi}_0} (Y - \hat{B}_1) \overline{Z}_{k_0}^T \right]_{i_1} \times \\ \prod_{s=2}^{t-1} \left( \frac{R_0}{\widehat{\pi}_0} \overline{Z}_{k_0} \overline{Z}_{k_0}^T - I \right)_{i_s} \left( \frac{R_0}{\widehat{\pi}_0} - 1 \right)_{i_t} \overline{Z}_{k_0, i_t} \end{array} \right\} \\ + \left\{ \begin{array}{l} \left[ \frac{R_1}{\widehat{\pi}_1} \frac{R_0}{\widehat{\pi}_0} (Y - \hat{B}_1) + \frac{R_0}{\widehat{\pi}_0} (\hat{B}_1 - \hat{B}_0) \right]_{i_1} \overline{Z}_{k_0, i_1}^T \times \\ \prod_{s=2}^{j-1} \left( \frac{R_0}{\widehat{\pi}_0} \overline{Z}_{k_0} \overline{Z}_{k_0}^T - I \right)_{i_s} \left( \frac{R_0}{\widehat{\pi}_0} - 1 \right)_{i_j} \overline{Z}_{k_0, i_j} \end{array} \right\} \end{array} \right] \end{aligned}$$

**Proof.** This follows directly from theorem 46 and the fact that  $\tilde{\tau}_{l,t}^{(j)}(\hat{\theta}) = \tau_{l,t}^{(j)}(\hat{\theta}) = 0$

for  $t > 1$ . ■

Suppose the class of estimators  $\{\hat{\psi}_m(\mathbf{O}), m \geq 1\}$  are defined as

$$\begin{aligned}\hat{\psi}_m(\mathbf{O}) &\equiv \hat{\psi}_1(\mathbf{O}) + \sum_{j=2}^m \hat{\psi}_{j,j}(\mathbf{O}) \\ \hat{\psi}_1(\mathbf{O}) &\equiv \psi(\hat{\theta}) + \mathbb{IF}_{1, \tilde{\psi}_{k_0, k_1, m}(\theta)}(\hat{\theta}) \\ \hat{\psi}_{j,j}(\mathbf{O}) &\equiv \mathbb{IF}_{j, j, \tilde{\psi}_{k_0, k_1, m}(\theta)}(\hat{\theta})\end{aligned}$$

Then

$$\begin{aligned}E_\theta(\hat{\psi}_m) - \psi(\theta) &= \left\{ \begin{aligned} &E_\theta\left(\psi(\hat{\theta}) + \mathbb{IF}_{m, \tilde{\psi}_{k_0, k_1, m}(\theta)}(\hat{\theta}) - \tilde{\psi}_{k_0, k_1, m}(\theta)\right) \\ &+ \left(\tilde{\psi}_{k_0, k_1, m}(\theta) - E_\theta\left[\mathbb{IF}_{m, \tilde{\psi}^\dagger(\theta)}(\hat{\theta}) + \psi(\hat{\theta})\right]\right) \\ &+ \left[E_\theta\left(\mathbb{IF}_{m, \tilde{\psi}^\dagger}(\hat{\theta}) + \psi(\hat{\theta})\right) - \tilde{\psi}^\dagger(\theta)\right] + \left(\tilde{\psi}^\dagger(\theta) - \psi(\theta)\right) \end{aligned} \right\} \quad (55)\end{aligned}$$

The following theorem examines the bias and variance of  $\hat{\psi}_m$ . Define  $\delta P_t = \frac{\pi_t}{\hat{\pi}_t} - 1$ ,

$\delta B_t = B_t - \hat{B}_t$ , and  $\delta g_t = \frac{g_t}{\hat{g}_t} - 1$  for  $t = 0, 1$ . Let  $q_0 = \frac{\pi_0}{\hat{\pi}_0}$  and  $q_{01} = \frac{\pi_0 \pi_1}{\hat{\pi}_0 \hat{\pi}_1}$ .

**Theorem 49** *Suppose conditions (a1) – (c2) hold then*

$$E_\theta(\hat{\psi}_1|\hat{\theta}) - \psi(\theta) = E_\theta\left[\frac{R_0}{\hat{\pi}_0}\delta P_1\delta B_1 + \delta P_0\delta B_0\middle|\hat{\theta}\right]$$

and  $\forall m > 1$

$$E_\theta(\hat{\psi}_m|\hat{\theta}) - \psi(\theta) = BI_{m,1} + BI_{m,2}$$



where

$$\begin{aligned}
& BI_{m,1} (-1)^{m-1} \\
& = \left\{ \begin{aligned} & \left\{ \begin{aligned} & E_{\theta} \left\{ [q_0 \delta B_0] \bar{Z}_{k_0}^T \right\} E_{\theta} \left[ q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T \right]^{-1} \times \\ & \left[ E_{\theta} \left( q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T \right) - I \right]^{m-1} E_{\theta} [\bar{Z}_{k_0} \delta P_0] \end{aligned} \right\}_{-(EB_1^{(1)})} \\ & + \left\{ \begin{aligned} & E_{\theta} \left\{ [q_{01} \delta B_1] \bar{W}_{k_1}^T \right\} E_{\theta} \left[ q_{01} \bar{W}_{k_1} \bar{W}_{k_1}^T \right]^{-1} \times \\ & \left[ E_{\theta} \left( q_{01} \bar{W}_{k_1} \bar{W}_{k_1}^T \right) - I \right]^{m-1} E_{\theta} [\bar{W}_{k_1} q_0 \delta P_1] \end{aligned} \right\}_{-(EB_1^{(2)})} \\ & + \sum_{j=2}^{m-1} \left\{ \begin{aligned} & E_{\theta} [\bar{Z}_{k_0} \delta P_0^T] \left[ E_{\theta} \left( q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T - I \right) \right]^{j-2} \times \\ & E_{\theta} [q_{01} \delta B_1 \bar{Z}_{k_0} \bar{W}_{k_1}^T] E_{\theta} [q_{01} \bar{W}_{k_1} \bar{W}_{k_1}^T]^{-1} \times \\ & E_{\theta} \left( q_{01} \bar{W}_{k_1} \bar{W}_{k_1}^T - I \right)^{m-j} E_{\theta} [\bar{W}_{k_1} q_0 \delta P_1] \end{aligned} \right\}_{-(EB_{jj}^{(2)})} \\ & + E_{\theta} \left\{ [q_0 \delta P_1 \delta B_1] \bar{Z}_{k_0}^T \right\} \left[ E_{\theta} \left( q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T \right) - I \right]^{m-2} E_{\theta} [\bar{Z}_{k_0} \delta P_0]_{-(EB_{mm}^{(2)})} \end{aligned} \right\} \quad (56)
\end{aligned}$$

$$\begin{aligned}
& |BI_{m,1}| \\
& \leq \left[ \begin{aligned}
& \left\{ \begin{aligned}
& \left\| \widehat{f}_0 \right\|_{\infty} \left\| q_0^{-1/2} \right\|_{\infty} \|q_0\|_{\infty} \left\| \frac{\widehat{f}_0}{f_0} \right\|_{\infty}^{\frac{1}{2}} \times \\
& \left\| \delta g_0 \right\|_{\infty}^{m-1} \left\{ \int \left( b_0 - \widehat{b}_0 \right)^2 dL_0 \right\}^{1/2} \times \\
& \left\| \frac{f_0}{\pi_0 \widehat{\pi}_0} \right\|_{\infty} \left\{ \int (\pi_0 - \widehat{\pi}_0)^2 dL_0 \right\}^{1/2}
\end{aligned} \right\} \\
& + \left\{ \begin{aligned}
& \left\| \widehat{f}_1 \right\|_{\infty} \left\| q_{01}^{-1/2} \right\|_{\infty} \|q_{01}\|_{\infty} \left\| \frac{\widehat{f}_1}{f_1} \right\|_{\infty}^{\frac{1}{2}} \times \\
& \left\| \delta g_1 \right\|_{\infty}^{m-1} \left\{ \int \left( b_1 - \widehat{b}_1 \right)^2 dL_0 dL_1 \right\}^{1/2} \\
& \times \left\| \frac{\pi_0 f_1}{\widehat{\pi}_0 \pi_1 \widehat{\pi}_1} \right\|_{\infty} \left\{ \int (\pi_1 - \widehat{\pi}_1)^2 dL_0 dL_1 \right\}^{1/2}
\end{aligned} \right\} \\
& + \sum_{j=2}^{m-1} \left\{ \begin{aligned}
& \left\| \frac{f_1}{\widehat{f}_1} \right\|_{\infty} \left\| \frac{f_0^2}{\widehat{f}_0 \widehat{\pi}_0^2} \right\|_{\infty} \left\| \frac{\widehat{f}_1}{f_1} \right\|_{\infty} \left\| \frac{\pi_0 f_1}{\widehat{\pi}_0 \pi_1 \widehat{\pi}_1} \right\|_{\infty} \\
& \|q_{01}\|_{\infty} \|\delta B_1\|_{\infty} \|\delta g_0\|_{\infty}^{j-2} \left\{ \int (\pi_0 - \widehat{\pi}_0)^2 dL_0 \right\}^{1/2} \times \\
& \left\| q_{01}^{-1/2} \right\|_{\infty} \|\delta g_1\|_{\infty}^{m-j} \left\{ \int (\pi_1 - \widehat{\pi}_1)^2 dL_0 dL_1 \right\}^{1/2} \\
& + \left\{ \begin{aligned}
& \left\| \frac{f_1}{\widehat{f}_1 \widehat{\pi}_1^2} q_0 \right\|_{\infty} \|\delta B_1\|_{\infty} \left\{ \int (\pi_1 - \widehat{\pi}_1)^2 dL_0 dL_1 \right\}^{1/2} \\
& \times \left\| \frac{f_0^2}{\widehat{f}_0 \widehat{\pi}_0^2} \right\|_{\infty} \|\delta g_0\|_{\infty}^{m-2} \left\{ \int (\pi_0 - \widehat{\pi}_0)^2 dL_0 \right\}^{1/2}
\end{aligned} \right\}
\end{aligned} \right\} \\
& \left. \right] \\
& = O_p \left( \max \left[ \begin{aligned}
& \left( \frac{\log n}{n} \right)^{\frac{(m-1)\beta_{g_0}}{d_0+2\beta_{g_0}}} n^{-\left( \frac{\beta_{b_0}}{d_0+2\beta_{b_0}} + \frac{\beta_{\pi_0}}{d_0+2\beta_{\pi_0}} \right)}, \\
& \left( \frac{\log n}{n} \right)^{\frac{(m-1)\beta_{g_1}}{d_1+2\beta_{g_1}}} n^{-\left( \frac{\beta_{b_1}}{d_1+2\beta_{b_1}} + \frac{\beta_{\pi_1}}{d_1+2\beta_{\pi_1}} \right)}, \\
& \sum_{j=2}^m \left( \frac{\log n}{n} \right)^{\frac{\beta_{b_1}}{d_1+2\beta_{b_1}} + \frac{(j-2)\beta_{g_0}}{d_0+2\beta_{g_0}} + \frac{(m-j)\beta_{g_1}}{d_1+2\beta_{g_1}}} n^{-\frac{\beta_{\pi_0}}{d_0+2\beta_{\pi_0}} - \frac{\beta_{\pi_1}}{d_1+2\beta_{\pi_1}}}
\end{aligned} \right] \right)
\end{aligned}
\tag{57}$$

$$\tag{58}$$

and

$$\begin{aligned}
& BI_{m,2} \\
&= \left\{ \begin{aligned} & E_{\theta} \left( \Pi_{\theta}^{\perp} \left( q_0^{1/2} \delta B_0 | \left( q_0^{1/2} \overline{Z}_{k_0} \right) \right) \Pi_{\theta}^{\perp} \left( q_0^{-1/2} \delta P_0 | \left( q_0^{1/2} \overline{Z}_{k_0} \right) \right) \right) \\ & + E_{\theta} \left( \Pi_{\theta}^{\perp} \left( q_{01}^{1/2} \delta B_1 | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right) \Pi_{\theta}^{\perp} \left( q_0 q_{01}^{-1/2} \delta P_1 | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right) \right) \end{aligned} \right\} + I(m > 2) \times \\
& \left\{ \begin{aligned} & -E_{\theta} \left[ \begin{aligned} & \Pi_{\theta}^{\perp} \left[ \left( \frac{\pi_1^{1/2}}{\pi_1^{1/2}} \delta B_1 \Pi_{\theta} \left( q_0^{1/2} \delta P_0 | \left( q_0^{1/2} \overline{Z}_{k_0} \right) \right) \right) | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right] \\ & \times \Pi_{\theta}^{\perp} \left[ q_0 q_{01}^{-1/2} \delta P_1 | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right] \end{aligned} \right]_{-(TB(m,2))} \\ & + (-1)^m E_{\theta} \left[ \begin{aligned} & \Pi_{\theta}^{\perp} \left[ \left( \begin{aligned} & q_{01}^{1/2} \delta B_1 E_{\theta} \left[ \delta P_0 \overline{Z}_{k_0}^T \right] E_{\theta} \left[ q_0 \overline{Z}_{k_0} \overline{Z}_{k_0}^T \right]^{-1} \right) | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right] \\ & \times \left( E_{\theta} \left[ q_0 \overline{Z}_{k_0} \overline{Z}_{k_0}^T - I \right] \right)^{m-2} \overline{Z}_{k_0} \end{aligned} \right) \right] \\ & \times \Pi_{\theta}^{\perp} \left[ q_0 q_{01}^{-1/2} \delta P_1 | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right] \end{aligned} \right]_{-(TB(m,3))} \end{aligned} \right\} \quad (59)
\end{aligned}$$

$$\begin{aligned}
& |BI_{m,2}| \\
&= O_p \left( \max \left[ \begin{aligned} & k_0^{-(\beta_{b_0} + \beta_{\pi_0})/d_0}, k_1^{-(\beta_{b_1} + \beta_{\pi_1})/d_1}, k_1^{-\beta_{\pi_1}/d_1} k_0^{-\beta_{\pi_0}/d_0} \left( \frac{\log n}{n} \right)^{\frac{\beta_{b_1}}{d_1 + \beta_{b_1}}} \\ & k_1^{-(\min(\beta_{\pi_0}, \beta_{b_1}) + \beta_{\pi_1})/d_1}, \left( \frac{\log n}{n} \right)^{-\frac{\beta_{b_1}}{d_1 + \beta_{b_1}} - \frac{(m-2)\beta_{g_0}}{d_0 + \beta_{g_0}}} n^{-\frac{\beta_{\pi_0}}{d_0 + \beta_{\pi_0}}} k_1^{-\beta_{\pi_1}/d_1} \end{aligned} \right] \right) \quad (60)
\end{aligned}$$

Moreover,

$$\text{var} \left( \widehat{\psi}_m | \widehat{\theta} \right) = O_P \left( \frac{1}{n} \max \left\{ 1, \frac{1}{n^m} \max \left( k_0^{m-1}, k_0^{m-2} k_1, \dots, k_0 k_1^{m-2}, k_1^{m-1} \right) \right\} \right)$$

The optimal choice of  $k_0$  and  $k_1$  in this class of higher order estimators, will depend on the size of the effective smoothness exponents  $\frac{\beta_{b_1}}{d_1}, \frac{\beta_{\pi_1}}{d_1}$  and  $\frac{\beta_{g_1}}{d_1}$  relative to

the size of the exponents  $\frac{\beta_{b_0}}{d_0}, \frac{\beta_{\pi_0}}{d_0}$  and  $\frac{\beta_{g_0}}{d_0}$ . A general prescription for finding an optimal estimator in our class is to choose a pair  $(k_{0,opt}(m_{opt}), k_{1,opt}(m_{opt}))$  that minimizes the maximum asymptotic MSE over the model among the candidate  $\hat{\psi}_m = \hat{\psi}_{m,(k_0,k_1)}$ . Here, the estimator  $\hat{\psi}_{m,(k_{0,opt}(m), k_{1,opt}(m))}$  uses the pair  $(k_{0,opt}(m), k_{1,opt}(m))$  of  $m$  that equates the order of the variance to the order of the maximum between the squared truncation and estimation biases (which are given in the theorem)

### 7.2.3 Three-occasion Monotone Missing Data

Theorem 46 can be applied in a nested fashion to the estimation of general functionals in nonparametric models with monotone missingness (that is with an arbitrary number of occasions). Building on the two-occasion case,

$$O_i = (L_0, R_0, R_0 L_1, R_1, R_0 R_1 L_2, R_2, R_0 R_1 R_2 Y),$$

where  $R_2$  is the missing indicator for the third occasion. We also write  $\pi_2 = \Pr(R_2 = 1 | R_0 = R_1 = 1, L_0, L_1, L_2)$ , and  $B_2 = E(Y | L_0, L_1, L_2)$ . Let

$$\bar{V}_{k_2} \equiv \hat{E}(\Phi_{k_2}(L_0, L_1, L_2) \Phi_{k_2}^T(L_0, L_1, L_2))^{-1/2} \Phi_{k_2}(L_0, L_1, L_2)$$

and  $\Phi_{k_2}(L_0, L_1, L_2)$  is a  $k_2$ -dimensional vector of tensor product basis for functions of  $(L_0, L_1, L_2)$  with finite variance. The truncated parameter  $\tilde{\psi}_{k_0, k_1, k_2}^{(3)}(\theta)$  is given by :

$$\tilde{\psi}_{k_0, k_1, k_2}^{(3)}(\theta) = \psi(\hat{\theta}) + \tilde{\tau}_1^{(3)}(\theta) + \sum_{j=2}^m \tilde{\psi}_{j,j}^{(3)}(\theta)$$

where for all  $1 \leq s \leq k_1, 1 \leq t \leq k_0$ ,

$$\begin{aligned}
\tilde{\pi}_2^{(3)-1} &= \left\{ \hat{\pi}_2^{-1} \begin{pmatrix} 1 - \bar{V}_{k_2}^T E_\theta \left( \frac{R_2}{\hat{\pi}_2} \frac{R_1}{\hat{\pi}_1} \frac{R_0}{\hat{\pi}_0} \bar{V}_{k_2} \bar{V}_{k_2}^T \right)^{-1} \\ \times E_\theta \left[ \frac{R_0}{\hat{\pi}_0} \frac{R_1}{\hat{\pi}_1} \left( \frac{R_2}{\hat{\pi}_2} - 1 \right) \bar{V}_{k_2} \right] \end{pmatrix} \right\} \\
\tilde{B}_2^{(3)} &= \left\{ \hat{B}_2 + \bar{V}_{k_2}^T E_\theta \left( \frac{R_2}{\hat{\pi}_2} \frac{R_1}{\hat{\pi}_1} \frac{R_0}{\hat{\pi}_0} \bar{V}_{k_2} \bar{V}_{k_2}^T \right)^{-1} \right. \\
&\quad \left. \times E_\theta \left( \frac{R_2}{\hat{\pi}_2} \frac{R_1}{\hat{\pi}_1} \frac{R_0}{\hat{\pi}_0} \left( Y - \hat{B}_2 \right) \bar{V}_{k_2} \right) \right\} \\
\widetilde{W_s B_2}^{(3)} &= \left\{ \hat{B}_2 W_s + \bar{V}_{k_2}^T E_\theta \left( \frac{R_2}{\hat{\pi}_2} \frac{R_1}{\hat{\pi}_1} \frac{R_0}{\hat{\pi}_0} \bar{V}_{k_2} \bar{V}_{k_2}^T \right)^{-1} \right. \\
&\quad \left. \times E_\theta \left( \frac{R_2}{\hat{\pi}_2} \frac{R_1}{\hat{\pi}_1} \frac{R_0}{\hat{\pi}_0} \left( Y - \hat{B}_2 \right) W_s \bar{V}_{k_2} \right) \right\} \\
\widetilde{Z_t W_s B_2}^{(3)} &= \left\{ \hat{B}_2 Z_t W_s + \bar{V}_{k_2}^T E_\theta \left( \frac{R_2}{\hat{\pi}_2} \frac{R_1}{\hat{\pi}_1} \frac{R_0}{\hat{\pi}_0} \bar{V}_{k_2} \bar{V}_{k_2}^T \right)^{-1} \right. \\
&\quad \left. \times E_\theta \left( \frac{R_2}{\hat{\pi}_2} \frac{R_1}{\hat{\pi}_1} \frac{R_0}{\hat{\pi}_0} \left( Y - \hat{B}_2 \right) Z_t W_s \bar{V}_{k_2} \right) \right\} \\
\widetilde{Z_t B_2}^{(3)} &= \left\{ \hat{B}_2 Z_t + \bar{V}_{k_2}^T E_\theta \left( \frac{R_2}{\hat{\pi}_2} \frac{R_1}{\hat{\pi}_1} \frac{R_0}{\hat{\pi}_0} \bar{V}_{k_2} \bar{V}_{k_2}^T \right)^{-1} \right. \\
&\quad \left. \times E_\theta \left( \frac{R_2}{\hat{\pi}_2} \frac{R_1}{\hat{\pi}_1} \frac{R_0}{\hat{\pi}_0} \left( Y - \hat{B}_2 \right) Z_t \bar{V}_{k_2} \right) \right\}
\end{aligned}$$

and

$$\tilde{\tau}_1^{(3)}(\theta) = \left\{ \begin{aligned} &E \left( \frac{R_2}{\hat{\pi}_2^{(3)}} \frac{R_1}{\hat{\pi}_1} \frac{R_0}{\hat{\pi}_0} \left( Y - \tilde{B}_2^{(3)} \right) + \frac{R_1}{\hat{\pi}_1} \frac{R_0}{\hat{\pi}_0} \left( \tilde{B}_2^{(3)} - \hat{B}_1 \right) \right) \\ &E \left( \frac{R_0}{\hat{\pi}_0} \left( \hat{B}_1 - \hat{B}_0 \right) + \hat{B}_0 - \psi(\hat{\theta}) \right) \end{aligned} \right\}$$

$$\tilde{\psi}_{j,j}^{(3)}(\theta)$$

$$= (-1)^{j-1} \left[ \begin{aligned} & \left\{ (A_s)_{1 \times k_1} \left[ E \left( \frac{R_0}{\hat{\pi}_0} \frac{R_1}{\hat{\pi}_1} \overline{W}_{k_1} \overline{W}_{k_1}^T - I \right) \right]^{j-2} E \left[ \overline{W}_{k_1} \frac{R_0}{\hat{\pi}_0} \left( \frac{R_1}{\hat{\pi}_1} - 1 \right) \right] \right\} \\ & + \sum_{q=2}^{j-1} \left\{ \begin{aligned} & E \left( \left( \frac{R_0}{\hat{\pi}_0} \left( \frac{R_1}{\hat{\pi}_1} - 1 \right) \right) \overline{W}_{k_1}^T \right) \left[ E \left( \frac{R_1}{\hat{\pi}_1} \frac{R_0}{\hat{\pi}_0} \overline{W}_{k_1} \overline{W}_{k_1}^T - I \right) \right]^{j-q-1} \\ & \times (B_{s,t})_{k_1 \times k_0} \times \\ & \left[ E \left( \frac{R_0}{\hat{\pi}_0} \overline{Z}_{k_0} \overline{Z}_{k_0}^T - I \right) \right]^{q-2} E \left[ \left( \frac{R_0}{\hat{\pi}_0} - 1 \right) \overline{Z}_{k_0} \right] \end{aligned} \right\} \\ & + \left\{ (C_t)_{1 \times k_0} \times \left[ E \left( \frac{R_0}{\hat{\pi}_0} \overline{Z}_{k_0} \overline{Z}_{k_0}^T - I \right) \right]^{j-2} E \left[ \left( \frac{R_0}{\hat{\pi}_0} - 1 \right) \overline{Z}_{k_0} \right] \right\} \end{aligned} \right]$$

where

$$\begin{aligned} A_s &= E \left( \frac{R_2}{\hat{\pi}_2^{(3)}} \frac{R_0}{\hat{\pi}_0} \frac{R_1}{\hat{\pi}_1} \left( Y W_s - \widetilde{W_s B_2^{(3)}} \right) + \frac{R_0}{\hat{\pi}_0} \frac{R_1}{\hat{\pi}_1} \left( \widetilde{W_s B_2^{(3)}} - \hat{B}_1 W_s \right) \right) \\ B_{s,t} &= E \left( \frac{R_2}{\hat{\pi}_2^{(3)}} \frac{R_0}{\hat{\pi}_0} \frac{R_1}{\hat{\pi}_1} \left( Y W_s Z_t - \widetilde{Z_t W_s B_2^{(3)}} \right) + \frac{R_0}{\hat{\pi}_0} \frac{R_1}{\hat{\pi}_1} \left( \widetilde{Z_t W_s B_2^{(3)}} - \hat{B}_1 W_s Z_t \right) \right) \\ C_t &= E \left( \frac{R_2}{\hat{\pi}_2^{(3)}} \frac{R_0}{\hat{\pi}_0} \frac{R_1}{\hat{\pi}_1} \left( Y Z_t - \widetilde{Z_t B_2^{(3)}} \right) + \frac{R_0}{\hat{\pi}_0} \frac{R_1}{\hat{\pi}_1} \left( \widetilde{Z_t B_2^{(3)}} - \hat{B}_1 Z_t \right) \right) \\ &\quad + E \left( \frac{R_0}{\hat{\pi}_0} \left( \hat{B}_1 - \hat{B}_0 \right) Z_t \right) \end{aligned}$$

We only give out the final expression for  $IF_{r,r,\tilde{\psi}_{k_0,k_1,k_2}^{(3)}}(\hat{\theta})$  without any technical details; bias and variance properties of this estimator will be published elsewhere.

$$\begin{aligned}
& (-1)^{r-1} IF_{r,r,\tilde{\psi}_{k_0,k_1,k_2}^{(3)}(\theta)}(\hat{\theta}) \\
&= \left[ \frac{R_0 R_1 R_2}{\hat{\pi}_0 \hat{\pi}_1 \hat{\pi}_2} (Y - \hat{B}_2) \bar{V}_{k_2}^T \right] \prod_{q=2}^{r-1} \left( \frac{R_0 R_1 R_2}{\hat{\pi}_0 \hat{\pi}_1 \hat{\pi}_2} \bar{V}_{k_2} \bar{V}_{k_2}^T - I \right)_{i_q} \left[ \bar{V}_{k_2} \frac{R_0 R_1}{\hat{\pi}_0 \hat{\pi}_1} \left( \frac{R_2}{\hat{\pi}_2} - 1 \right) \right]_{i_r} \\
&+ \sum_{m=2}^{r-1} \left\{ + \sum_{j=2}^{m-1} \left[ \begin{aligned} & \left[ \frac{R_0}{\hat{\pi}_0} \left( \frac{R_1}{\hat{\pi}_1} - 1 \right) \right]_{i_1} \bar{W}_{k_1}^T \prod_{q=2}^{m-1} \left( \frac{R_0 R_1}{\hat{\pi}_0 \hat{\pi}_1} \bar{W}_{k_1} \bar{W}_{k_1}^T - I \right)_{i_q} \times \\ & \left[ \frac{R_0 R_1 R_2}{\hat{\pi}_0 \hat{\pi}_1 \hat{\pi}_2} (Y - \hat{B}_2) \bar{W}_{k_1} \bar{V}_{k_2}^T \right]_{i_m} \prod_{s=m+1}^{r-1} \left( \frac{R_0 R_1 R_2}{\hat{\pi}_0 \hat{\pi}_1 \hat{\pi}_2} \bar{V}_{k_2} \bar{V}_{k_2}^T - I \right)_{i_s} \times \\ & \left[ \bar{V}_{k_2} \frac{R_0 R_1}{\hat{\pi}_0 \hat{\pi}_1} \left( \frac{R_2}{\hat{\pi}_2} - 1 \right) \right]_{i_r} \\ & \left[ \left( \frac{R_0}{\hat{\pi}_0} - 1 \right)_{i_1} \bar{Z}_{k_0,i_1}^T \prod_{s=2}^{j-1} \left( \frac{R_0}{\hat{\pi}_0} \bar{Z}_{k_0} \bar{Z}_{k_0}^T - I \right)_{i_s} \left[ \frac{R_0 R_1 R_2}{\hat{\pi}_0 \hat{\pi}_1 \hat{\pi}_2} (Y - \hat{B}_2) \bar{Z}_{k_0} \bar{V}_{k_2}^T \right]_{i_j} \right. \\ & \times \prod_{q=j+1}^{r+j-m-1} \left( \frac{R_0 R_1 R_2}{\hat{\pi}_0 \hat{\pi}_1 \hat{\pi}_2} \bar{V}_{k_2} \bar{V}_{k_2}^T - I \right)_{i_q} \left[ \bar{V}_{k_2} \frac{R_0 R_1}{\hat{\pi}_0 \hat{\pi}_1} \left( \frac{R_2}{\hat{\pi}_2} - 1 \right) \bar{W}_{k_1}^T \right]_{i_{r+j-m}} \times \\ & \left. \prod_{q=r+j-m+1}^{r-1} \left( \frac{R_0 R_1}{\hat{\pi}_0 \hat{\pi}_1} \bar{W}_{k_1} \bar{W}_{k_1}^T - I \right)_{i_q} \left[ \bar{W}_{k_1} \frac{R_0}{\hat{\pi}_0} \left( \frac{R_1}{\hat{\pi}_1} - 1 \right) \right]_{i_r} \right] \\ & + \left( \frac{R_0}{\hat{\pi}_0} - 1 \right)_{i_1} \bar{Z}_{k_0,i_1}^T \prod_{s=2}^{m-1} \left( \frac{R_0}{\hat{\pi}_0} \bar{Z}_{k_0} \bar{Z}_{k_0}^T - I \right)_{i_s} \left[ \bar{Z}_{k_0} \frac{R_0 R_1 R_2}{\hat{\pi}_0 \hat{\pi}_1 \hat{\pi}_2} (Y - \hat{B}_2) \bar{V}_{k_2}^T \right]_{i_m} \times \\ & \prod_{q=m+1}^{r-1} \left( \frac{R_0 R_1 R_2}{\hat{\pi}_0 \hat{\pi}_1 \hat{\pi}_2} \bar{V}_{k_2} \bar{V}_{k_2}^T - I \right) \left[ \bar{V}_{k_2} \frac{R_0 R_1}{\hat{\pi}_0 \hat{\pi}_1} \left( \frac{R_2}{\hat{\pi}_2} - 1 \right) \right]_{i_r} \end{aligned} \right\}
\end{aligned}$$

$$\begin{aligned}
& + \left\{ \begin{aligned} & \left[ \begin{aligned} & \frac{R_0 R_1 R_2}{\hat{\pi}_0 \hat{\pi}_1 \hat{\pi}_2} (Y - \hat{B}_2) + \\ & \frac{R_0 R_1}{\hat{\pi}_0 \hat{\pi}_1} (\hat{B}_2 - \hat{B}_1) \end{aligned} \right]_{i_1} \overline{W}_{k_1, i_1}^T \prod_{q=2}^{r-1} \left( \frac{R_0 R_1}{\hat{\pi}_0 \hat{\pi}_1} \overline{W}_{k_1} \overline{W}_{k_1}^T - I \right)_{i_q} \left[ \overline{W}_{k_1} \frac{R_0}{\hat{\pi}_0} \left( \frac{R_1}{\hat{\pi}_1} - 1 \right) \right]_{i_r} \\ & + \sum_{j=2}^{r-1} \left( \frac{R_0}{\hat{\pi}_0} - 1 \right)_{i_1} \overline{Z}_{k_0, i_1}^T \prod_{s=2}^{j-1} \left( \frac{R_0}{\hat{\pi}_0} \overline{Z}_{k_0} \overline{Z}_{k_0}^T - I \right)_{i_s} \left[ \begin{aligned} & \left( \frac{R_0 R_1 R_2}{\hat{\pi}_0 \hat{\pi}_1 \hat{\pi}_2} (Y - \hat{B}_2) \right) \\ & + \frac{R_0 R_1}{\hat{\pi}_0 \hat{\pi}_1} (\hat{B}_2 - \hat{B}_1) \end{aligned} \right] \overline{Z}_{k_0} \overline{W}_{k_1}^T \Big]_{i_j} \times \\ & \prod_{q=j+1}^{r-1} \left( \frac{R_0 R_1}{\hat{\pi}_0 \hat{\pi}_1} \overline{W}_{k_1} \overline{W}_{k_1}^T - I \right)_{i_q} \left[ \overline{W}_{k_1} \frac{R_0}{\hat{\pi}_0} \left( \frac{R_1}{\hat{\pi}_1} - 1 \right) \right]_{i_r} \\ & + \left[ \begin{aligned} & \left( \frac{R_0 R_1 R_2}{\hat{\pi}_0 \hat{\pi}_1 \hat{\pi}_2} (Y - \hat{B}_2) \right) \\ & + \frac{R_0 R_1}{\hat{\pi}_0 \hat{\pi}_1} (\hat{B}_2 - \hat{B}_1) \\ & + \frac{R_0}{\hat{\pi}_0} (\hat{B}_1 - \hat{B}_0) \end{aligned} \right]_{i_1} \overline{Z}_{k_0, i_1}^T \prod_{s=2}^{r-1} \left( \frac{R_0}{\hat{\pi}_0} \overline{Z}_{k_0} \overline{Z}_{k_0}^T - I \right)_{i_s} \overline{Z}_{k_0, i_r} \left( \frac{R_0}{\hat{\pi}_0} - 1 \right)_{i_r} \end{aligned} \right\}
\end{aligned}$$

## References

- Arellano M. (2003) ” *Panel Data Econometrics*”. Oxford University Press: Advanced Texts in Econometrics.
- Bhattacharyya A. (1947) ”On some analogues of the amount of information and their use in statistical estimation II-III” *Sankhyā*, Vol.8, 3, 201–218.
- Bickel P. , Klassen C., Ritov Y., Wellner J. (1993) ” *Efficient and Adaptive Estimation for Semiparametric Models*”. Springer.
- Bickel P., and Ritov Y. (2003) ”Nonparametric estimators which can be ”plugged-in”. *Annals of Statist.* 31(4), 1033–53.



Birge L., Massart P.(1995) "Estimation of Integral Functionals of a Density". *Annals of Statistics*. 23(1), 11-29.

He, X. and Shao, Q.M. (2000). "On parameters of increasing dimension". *Journal of Multivariate Analysis*, 73(1), 120-135.

Klassen C. (1987) "Consistent Estimation of the Influence Function of Locally Asymptotically Linear Estimators". *Annals of Statistics*. 15(4), 1548-62.

Lee A.J. (1990) "*U-Statistics: Theory and Practice*". Marcel Dekker, New York.

Lindsay R, Waterman B. (1996) "Projected Score Methods for Approximating Conditional Scores". *Biometrika* 83(1):1-13.

Li L., Tchetgen E., van der Vaart AW, and Robins JM. (2006) "Robust Inference with Higher Order Inference Functions: Part II." 2005 *JSM Proceedings*. American Statistical Associations. 2558-2565.

Mallat SG. (1998) "*A Wavelet Tour of Signal Processing*". Academic Press.

Newey W., Hsieh F., and Robins JM. (2004) "Twicing kernels and a small bias property of semiparametric estimators" *Econometrica* 72, 947-962.

Pfanzagl J. (1990) "*Estimation in Semiparametric Models: Some Recent Developments*". Lecture Notes in Statistics 31, Springer-Verlag, Berlin 1985, 505 pages

- Portnoy S. (1988). "Asymptotic Behavior of Likelihood Methods for Exponential Families When the Number of Parameters Tends to Infinity". *Annals of Statistics*. 16, 356-366
- Pyke R., "Spacings (with discussion)", *J. Roy. Statis. Soc. B* 27 (1965), 395–449.
- Ritov Y., Bickel P. (1990) "Achieving information bounds in non- and semiparametric models". *Annals of Statistics*. 18, 925-938.
- Robins J., Ritov, Y (1997). "Toward a Curse-of-dimensionality Appropriate (CODA) Asymptotic Theory for Semiparametric Models". *Statistics in Medicine*. 16 285-319.
- Robins JM. (2004) "Optimal Structural Nested Models for Optimal Sequential Decisions". in DY Lin and P. Heagerty (Eds.) *Proceedings of the Second Seattle Symposium in Biostatistics*. New York Springer.
- Robins JM, Rotnitzky A. (2001).Comment on the Bickel and Kwon article, "Inference for semiparametric models: Some questions and an answer" *Statistica Sinica*, 11(4):920-936. ["On Double Robustness."]
- Robins J.M, Li L, Tchetgen Eric, van der Vaart AW (2007), "Asymptotic Normality of Degenerate U-statistics". Working paper.

Robins J.M, van der Vaart AW. (2006) " Adaptive Nonparametric Confidence Sets". *Annals of statistics*. 34(1):229-253.

Robins JM, Li L, Tchetgen E, van der Vaart A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. *Probability and Statistics: Essays in Honor of David A. Freedman* 2:335-421.

Small D., McLeish C.(1994) " *Hilbert Space Methods in Probability and Statistical Inference*". Wiley.

Tchetgen E., Li L., van der Vaart AW, and Robins JM. (2006) "Robust Inference with Higher Order Inference Functions: Part I." 2005 *JSM Proceedings*. American Statistical Associations. 2644-2651.

Tchetgen E., Li L., van der Vaart AW, and Robins JM. (2007) "Higher Order U-statistics Estimators for Longitudinal Missing Data and Causal Inference Models". working paper

van der Laan M and Dudoit S.(2003). Asymptotics of Cross-Validated Risk Estimation in Estimator Selection and Performance Assessment. Technical report.

van der Laan M, and Robins JM. (2003) " *Unified Methods for Censored Longitudinal Data and Causality*" Springer Series in Statistics.

Wang, L., Brown, L. D., Cai, T. Levine, M. (2006). "Effect of mean on variance function estimation in nonparametric regression". Technical Report .

Cai, T., Levine, M. Wang, L. (2006) "Variance function estimation in multivariate nonparametric regression". Technical Report.

van der Vaart, AW (1991). "On Differentiable Functionals". *Ann. Statist.* 19 178-204.

van der Vaart, AW (1998). " *Asymptotic Statistics*". Cambridge Series in Statistical and Probabilistic Mathematics.

## 8 Appendix

In the following, we assume all parametric submodels are sufficiently smooth and regular that expectation and differentiation operators commute as needed. We also define  $\mathbb{IF}_{1,1}$  to be  $\mathbb{IF}_1$ .

**Proof.** (Theorem 2) Define the bias function  $B_m [\theta^\dagger, \theta]$  of  $\mathbb{IF}_m (\theta)$  to be  $E_{\theta^\dagger} [\mathbb{IF}_m (\theta)]$ .

Define

$$B_{m, l_1^* \dots l_j^* l_{j+1} \dots l_s} [\theta, \theta] = \partial^s B_m \left[ \tilde{\theta}(\varsigma^*), \tilde{\theta}(\varsigma) \right] / \partial \varsigma_{l_1}^* \dots \partial \varsigma_{l_j}^* \partial \varsigma_{l_{j+1}} \dots \partial \varsigma_{l_s} \Big|_{\varsigma^* = \tilde{\theta}^{-1}\{\theta\}, \varsigma = \tilde{\theta}^{-1}\{\theta\}}$$

where we reserve  $*$  for differentiation with respect to the first argument of  $B_m [\cdot, \cdot]$ .

Thus for  $s \leq m$ ,

$$\psi_{\setminus l_1 \dots l_s}(\theta) = B_{m, l_1^* \dots l_s^*}[\theta, \theta]$$

To prove the theorem we will first need to show that:

$$B_{m, l_1^* \dots l_j^* l_{j+1}, \dots l_s}[\theta, \theta] = 0 \text{ for } m \geq s > j > 0 \quad (61)$$

To this end note that for  $j < m$ ,

$$\begin{aligned} \psi_{\setminus l_1 \dots l_{j+1}}(\theta) &= \partial \psi_{\setminus l_1 \dots l_j}(\theta) / \partial \varsigma_{l_{j+1}} = \partial B_{m, l_1^* \dots l_j^*}[\theta, \theta] / \partial \varsigma_{l_{j+1}} \\ &= B_{m, l_1^* \dots l_j^* l_{j+1}^*}[\theta, \theta] + B_{m, l_1^* \dots l_j^* l_{j+1}}[\theta, \theta] \\ &= \psi_{\setminus l_1 \dots l_{j+1}}(\theta) + B_{m, l_1^* \dots l_j^* l_{j+1}}[\theta, \theta] \end{aligned}$$

where the 2nd equality is by the definition of  $\mathbb{IF}_m(\theta)$ , the third is by the chain rule, and the fourth is again by the definition of  $\mathbb{IF}_m(\theta)$ . Hence  $B_{m, l_1^* \dots l_j^* l_{j+1}}[\theta, \theta] = 0$ .

Hence for  $j \leq m - 2$ ,

$$\begin{aligned} 0 &= \partial B_{m, l_1^* \dots l_j^* l_{j+1}}[\theta, \theta] / \partial \varsigma_{l_{j+2}} = B_{m, l_1^* \dots l_j^* l_{j+2}^* l_{j+1}}[\theta, \theta] + B_{m, l_1^* \dots l_j^* l_{j+1} l_{j+2}}[\theta, \theta] \\ &= 0 + B_{m, l_1^* \dots l_j^* l_{j+1} l_{j+2}}[\theta, \theta] \end{aligned}$$

where the last equality holds because we just proved  $B_{m, l_1^* \dots l_j^* l_{j+1}}[\theta, \theta] = 0$  for arbitrary indices. Iterating this argument proves (61). We complete the proof by induction on  $s$  for some  $s < m$ . Given a  $s = 1$  dimensional regular parametric submodel  $\tilde{\theta}(\varsigma)$ ,  $E_{\theta(\varsigma)}[\mathbb{IF}_m(\theta(\varsigma))] = 0$  by assumption. Hence, by regularity of the model,  $0 =$

$B_{m,l_1^*}[\theta, \theta] + B_{m,l_1}[\theta, \theta]$ . Therefore  $B_{m,l_1}[\theta, \theta] = -\psi_{\setminus l_1}(\theta)$ . Now suppose the theorem is true for  $s$ . Then

$$\begin{aligned} -\psi_{\setminus l_1 \dots l_{s+1}}(\theta) &= -\partial \psi_{\setminus l_1 \dots l_s}(\theta) / \partial \zeta_{l_{s+1}} = \partial B_{m,l_1 \dots l_s}[\theta, \theta] / \partial \zeta_{l_{s+1}} \\ &= B_{m,l_{s+1}^* l_1 \dots l_s}[\theta, \theta] + B_{m,l_1 \dots l_{s+1}}[\theta, \theta] = 0 + B_{m,l_1 \dots l_{s+1}}[\theta, \theta] \end{aligned}$$

where the second equality is by the induction assumption, the third by the chain rule, and the last by equation (61) ■

**Proof.** (Theorem 3) (1): Consider two influence functions  $\mathbb{IF}_m^{(1)}(\theta)$  and  $\mathbb{IF}_m^{(2)}(\theta)$  for  $\psi(\theta)$ . Then  $E_\theta \left[ \left\{ \mathbb{IF}_m^{(1)}(\theta) - \mathbb{IF}_m^{(2)}(\theta) \right\} \tilde{\mathbb{S}}_{s, \bar{l}_s}(\theta) \right] = \psi_{\setminus \bar{l}_s}(\theta) - \psi_{\setminus \bar{l}_s}(\theta) = 0$  for any score  $\tilde{\mathbb{S}}_{s, \bar{l}_s}(\theta)$ ,  $s \leq m$  and hence for any linear combination of scores. But, by definition, linear combinations of scores are dense in  $\Gamma_m(\theta)$ . Thus  $\mathbb{IF}_m^{(1)}(\theta)$  and  $\mathbb{IF}_m^{(2)}(\theta)$  have the same projection on  $\Gamma_m(\theta)$ . (2-3): Essentially immediate from the definitions. (4): For  $t \leq s$ ,  $\psi_{\setminus \bar{l}_t}(\theta) = E_\theta \left[ \mathbb{IF}_m(\theta) \tilde{\mathbb{S}}_{t, \bar{l}_t}(\theta) \right] = E_\theta \left[ \Pi_{m, \theta} [\mathbb{IF}_m(\theta) | \mathcal{U}_t(\theta)] \tilde{\mathbb{S}}_{t, \bar{l}_t}(\theta) \right]$  for any  $\tilde{\mathbb{S}}_{t, \bar{l}_t}(\theta)$ . (5.a): follows from (1). (5.b): follows from (4). Degeneracy of  $\mathbb{IF}_{mm}(\theta)$  follows at once from the fact that  $\mathbb{IF}_{mm}(\theta) \in \mathcal{U}_{m-1}(\theta)^\perp$  in  $\mathcal{U}_m(\theta)$ .

Proof of part (5.c) requires the following. ■

**Lemma 50** Suppose, for  $m \geq 1$ ,  $\mathbb{IF}_{m,m}(\theta)$  and  $if_{1, if_{m,m}}(O_{i_1}, \dots, O_{i_m}; \cdot)(O_{i_{m+1}}; \theta)$  exist w.p.1 for a kernel  $IF_{m,m}(\theta)$ . Then, (i)  $if_{1, if_{m,m}}(O_{i_1}, \dots, O_{i_m}; \cdot)(O_{i_{m+1}}; \theta) s_{l_t}(O_{i_{m+1}})$ ,  $-if_{m,m, \setminus l_t}(O_{i_1}, \dots, O_{i_m}; \theta)$ , and  $if_{m,m}(O_{i_1}, \dots, O_{i_m}; \cdot) s_{l_t}(O_{i_m})$  each have the same mean given  $O_{i_1}, \dots, O_{i_{m-1}}$ , (ii)  $E[if_{m,m, \setminus l_t}(O_{i_1}, \dots, O_{i_m}; \theta) | O_{i_1}, \dots, O_{i_{m-2}}] = 0$ ,

$$(iii) E_{\theta} \left[ if_{1, if_{m,m}}(O_{i_1}, \dots, O_{i_m}; \cdot) (O_{i_{m+1}}; \theta) | O_{i_1}, \dots, O_{i_{m-2}}, O_{i_{m+1}} \right] = 0, \text{ so}$$

$$\begin{aligned} & \Pi \left[ \mathbb{V} \left[ if_{1, if_{m,m}}(O_{i_1}, \dots, O_{i_m}; \cdot) (O_{i_{m+1}}; \theta) \right] | \mathcal{U}_m(\theta) \right] \\ &= \Pi \left[ \mathbb{V} \left[ if_{1, if_{m,m}}(O_{i_1}, \dots, O_{i_m}; \cdot) (O_{i_{m+1}}; \theta) \right] | \mathcal{U}_m(\theta) \cap \mathcal{U}_{m-2}^{\perp}(\theta) \right] \end{aligned}$$

and (iv)  $\mathbb{IF}_{m,m,\setminus l_t}(\theta)$  satisfies  $\Pi_{\theta} [\mathbb{IF}_{m,m,\setminus l_t}(\theta) | \mathcal{U}_{m-2}(\theta)] = 0$  and

$$\Pi_{\theta} [\mathbb{IF}_{m,m,\setminus l_t}(\theta) | \mathcal{U}_{m-1}(\theta)] = -\mathbb{V} \left[ m E_{\theta} \left[ IF_{m,m,\setminus l_t, \bar{i}_m}^{sym}(\theta) | O_{i_1}, \dots, O_{i_{m-1}} \right] \right].$$

**Proof.** (i): By  $IF_{m,m}(\theta)$  degenerate,

$E_{\theta} [IF_{m,m,\setminus l_t, \bar{i}_m}(\theta) | O_{i_1}, \dots, O_{i_{m-1}}] = -E_{\theta} [IF_{m,m,\bar{i}_m}(\theta) s_{l_t}(O_{i_m}) | O_{i_1}, \dots, O_{i_{m-1}}]$ . Further, by definition,

$$\begin{aligned} & E_{\theta} \left[ if_{1, if_{m,m}^{sym}}(O_{i_1}, \dots, O_{i_m}; \cdot) (O_{i_{m+1}}; \theta) s_{l_t}(O_{i_{m+1}}) | O_{i_1}, \dots, O_{i_m} \right] \\ &= E_{\theta} [IF_{m,m,\setminus l_t, \bar{i}_m}(\theta) | O_{i_1}, \dots, O_{i_m}]. \end{aligned}$$

(ii): By  $IF_{m,m}(\theta)$  degenerate  $0 = E_{\theta} [IF_{m,m,\bar{i}_m}(\theta) s_{l_t}(O_{i_m}) | O_{i_1}, \dots, O_{i_{m-2}}]$  w.p.1 and

so (ii) follows from (i).

(iii): (i) and (ii) imply

$$\begin{aligned} 0 &= E_{\theta} \left[ if_{1, if_{m,m}}(O_{i_1}, \dots, O_{i_m}; \cdot) (O_{i_{m+1}}; \theta) s_{l_t}(O_{i_{m+1}}) | O_{i_1}, \dots, O_{i_{m-2}} \right] \\ &= E_{\theta} \left[ \begin{aligned} & E_{\theta} \left\{ if_{1, if_{m,m}}(O_{i_1}, \dots, O_{i_m}; \cdot) (O_{i_{m+1}}; \theta) | (O_{i_{m+1}}, O_{i_1}, \dots, O_{i_{m-2}}) \right\} \\ & \times s_{l_t}(O_{i_{m+1}}) | O_{i_1}, \dots, O_{i_{m-2}} \end{aligned} \right] \end{aligned}$$

But, by  $s_{l_t}(O_{i_{m+1}})$  an arbitrary mean zero function,

$$\begin{aligned} & E_\theta \left\{ i f_{1, i f_{m-1, m}}(O_{i_1, \dots, O_{i_m}}; \cdot) (O_{i_{m+1}}; \theta) \mid (O_{i_{m+1}}, O_{i_1}, \dots, O_{i_{m-2}}) \right\} \\ & = E_\theta \left\{ i f_{1, i f_{m-1, m}}(O_{i_1, \dots, O_{i_m}}; \cdot) (O_{i_{m+1}}; \theta) \mid O_{i_1}, \dots, O_{i_{m-2}} \right\} = 0 \end{aligned}$$

(iv): By definition,  $\Pi_\theta [\mathbb{IF}_{m, m, \setminus l_t}(\theta) \mid \mathcal{U}_{m-1}(\theta)] = \mathbb{V} [\{I - d_{m, \theta}\} \{IF_{m, m, \setminus l_t, \bar{l}_m}(\theta)\}]$ .

The result follows by Eq. (4) and part (ii). ■

**Proof. Theorem 5c(ii):** Consider a  $m$ -dimensional parametric submodel  $f(O; \tilde{\theta}(\zeta)) = f(O; \theta) \{1 + \sum_{j=1}^m \zeta_j a_j(O)\}$ ,  $\zeta^T = (\zeta_1, \dots, \zeta_m)$ , with  $E_\theta[a_l(O)] = 0$ . Since this model is linear in the  $\zeta_j$ ,  $f_{/l_1 \dots l_m}(O_j; \theta) = 0$  for  $m > 1$ . Hence  $\tilde{\mathbb{S}}_{m, \bar{l}_m}(\theta)$  is degenerate of order  $m$ , i.e.,  $\tilde{\mathbb{S}}_{m, \bar{l}_m}(\theta) \in \mathcal{U}_{m-1}^\perp(\theta)$ . Since  $\mathbb{IF}_{m-1}(\theta)$  exists, on setting  $l_s = s$  for  $s = 1, \dots, m$ ,

$$\partial^{m-1} \psi(\tilde{\theta}(\zeta)) / \prod_{j=1}^{m-1} \partial \zeta_j |_{\zeta=0} \equiv \psi_{\setminus \bar{l}_{m-1}}(\theta) = E_\theta [\mathbb{IF}_{m-1}(\theta) \tilde{\mathbb{S}}_{m-1, \bar{l}_{m-1}}(\theta)].$$

Differentiating the last display with respect to  $\zeta_m$  and evaluating at  $\zeta = 0$ , we obtain

$$\begin{aligned} \psi_{\setminus \bar{l}_m}(\theta) &= E_\theta [\mathbb{IF}_{m-1}(\theta) \tilde{\mathbb{S}}_{m, \bar{l}_m}(\theta)] + E_\theta [\mathbb{IF}_{m-1, \setminus l_m}(\theta) \tilde{\mathbb{S}}_{m-1, \bar{l}_{m-1}}(\theta)] \\ &= E_\theta [\mathbb{IF}_{m-1, \setminus l_m}(\theta) \tilde{\mathbb{S}}_{m-1, \bar{l}_{m-1}}(\theta)] \end{aligned}$$

Now  $E_\theta [\mathbb{IF}_{m-1, \setminus l_m}(\theta) \tilde{\mathbb{S}}_{m-1, \bar{l}_{m-1}}(\theta)]$   
 $= E_\theta [\mathbb{IF}_{m-2, \setminus l_m}(\theta) \tilde{\mathbb{S}}_{m-1, \bar{l}_{m-1}}(\theta)] + E_\theta [\mathbb{IF}_{m-1, m-1, \setminus l_m}(\theta) \tilde{\mathbb{S}}_{m-1, \bar{l}_{m-1}}(\theta)]$ . Setting  $s_{l_r}(O_{i_r}, \theta) = a_r(O_{i_r})$ ,  $\tilde{\mathbb{S}}_{m-1, \bar{l}_{m-1}}(\theta) = \sum_{i_1 \neq \dots \neq i_{m-1}} \prod_{r=1}^{m-1} a_r(O_{i_r}, \theta)$  is degenerate of order  $m-1$  so

$$\begin{aligned} & E_\theta [\mathbb{IF}_{m-1, m-1, \setminus l_m}(\theta) \tilde{\mathbb{S}}_{m-1, \bar{l}_{m-1}}(\theta)] \\ &= (m-1)! E_\theta \left( [i f_{m-1, m-1, \setminus l_m}^{sym}(O_{i_1}, \dots, O_{i_{m-1}}; \theta)] \prod_{r=1}^{m-1} a_r(O_{i_r}, \theta) \right) \end{aligned}$$



and  $E_\theta \left[ \mathbb{IF}_{m-2, \setminus l_m}(\theta) \tilde{\mathbb{S}}_{m-1, \bar{l}_{m-1}}(\theta) \right] = 0$ . Hence

$$\psi_{\setminus \bar{l}_m}(\theta) = (m-1)! E_\theta \left( i f_{m-1, m-1, \setminus l_m}^{sym}(O_{i_1}, \dots, O_{i_{m-1}}; \theta) \prod_{r=1}^{m-1} a_r(O_{i_r}, \theta) \right)$$

Now, by the assumed existence of  $\mathbb{IF}_m(\theta)$ , we also have  $\psi_{\setminus \bar{l}_m}(\theta) = E_\theta \left[ \mathbb{IF}_m(\theta) \tilde{\mathbb{S}}_m(\theta) \right] = m! E_\theta \left( i f_{m, m}^{sym}(O_{i_1}, \dots, O_{i_m}; \theta) \prod_{r=1}^m a_r(O_{i_r}, \theta) \right)$ . It follows that, for any choice of  $m-1$  mean zero functions  $a_r(O)$  under  $\theta$ ,

$$\begin{aligned} 0 &= E_\theta \left( \left\{ \begin{array}{c} i f_{m-1, m-1, \setminus l_m}^{sym}(O_{i_1}, \dots, O_{i_{m-1}}; \theta) \\ -m E_\theta [i f_{m, m}^{sym}(O_{i_1}, \dots, O_{i_m}; \theta) a_m(O_{i_m}, \theta) | O_{i_1}, \dots, O_{i_{m-1}}] \end{array} \right\} \times \prod_{r=1}^{m-1} a_r(O_{i_r}, \theta) \right) \\ &= E_\theta \left( r(O_{i_1}, \dots, O_{i_{m-1}}; \theta) \prod_{r=1}^{m-1} a_r(O_{i_r}, \theta) \right) \end{aligned}$$

where

$$\begin{aligned} &r(O_{i_1}, \dots, O_{i_{m-1}}; \theta) \\ &\equiv d_{m-1, \theta} [i f_{m-1, m-1, \setminus l_m}^{sym}(O_{i_1}, \dots, O_{i_{m-1}}; \theta)] - m E_\theta [i f_{m, m}^{sym}(O_{i_1}, \dots, O_{i_m}; \theta) a_m(O_{i_m}, \theta) | O_{i_1}, \dots, O_{i_{m-1}}] \end{aligned}$$

The last equality follows from  $i f_{m-1, m-1, \setminus l_m}^{sym}(O_{i_1}, \dots, O_{i_{m-1}}; \theta) - d_{m-1, \theta} [i f_{m-1, m-1, \setminus l_m}^{sym}(O_{i_1}, \dots, O_{i_{m-1}}; \theta)]$  orthogonal to  $\prod_{r=1}^{m-1} a_r(O_{i_r}, \theta)$ . We conclude  $r(O_{i_1}, \dots, O_{i_{m-1}}; \theta) = 0$  with probability 1 because  $r(O_{i_1}, \dots, O_{i_{m-1}}; \theta)$  is a degenerate U-statistic kernel of order  $m-1$  and all degenerate U-statistics of order  $m-1$  have kernels that are the (possibly infinite) sum of products of  $m-1$  mean zero functions. It follows that, on a set  $\mathcal{O}_{m-1}$  which

has probability 1 under  $F^{(m-1)}(\cdot, \theta)$ ,

$$\begin{aligned}
& if_{m-1, m-1, \setminus l_m}^{sym}(o_{i_1}, \dots, o_{i_{m-1}}; \theta) \\
&= E_\theta \left[ \left\{ m \times if_{m, m}^{sym}(o_{i_1}, \dots, o_{i_{m-1}}, O_{i_m}; \theta) a_m(O_{i_m}, \theta) \right\} \right] \\
&+ \{I - d_{m-1, \theta}\} \left[ if_{m-1, m-1, \setminus l_m}^{sym}(o_{i_1}, \dots, o_{i_{m-1}}; \theta) \right] \\
&= E_\theta \left[ \left\{ \begin{aligned} & m \times if_{m, m}^{sym}(o_{i_1}, \dots, o_{i_{m-1}}, O; \theta) \\ & - \sum_{j=1}^{m-1} if_{m-1, m-1}^{sym}(o_{i_1}, \dots, o_{i_{j-1}}, O, o_{i_{j+1}}, \dots, o_{i_{m-1}}; \theta) \end{aligned} \right\} a_m(O, \theta) \right]
\end{aligned}$$

since, by parts (i) and (ii) of the lemma 50 and Eq. (4),

$$\begin{aligned}
& \{I - d_{m-1, \theta}\} \left[ if_{m-1, m-1, \setminus l_m}^{sym}(o_{i_1}, \dots, o_{i_{m-1}}; \theta) \right] \\
&= -E_\theta \left[ \sum_{j=1}^{m-1} if_{m-1, m-1}^{sym}(o_{i_1}, \dots, o_{i_{j-1}}, O, o_{i_{j+1}}, \dots, o_{i_{m-1}}; \theta) a_m(O, \theta) \right]
\end{aligned}$$

Here  $I$  is the identity operator. Now since the model  $f(O; \tilde{\theta}(\zeta)) = f(O; \theta) \{1 + \zeta_m a_m(O)\}$

with  $\zeta_s = 0$  for  $s < m$  has score  $a_m(O)$  and such scores are dense in the subspace of

$L_2(F(\cdot, \theta))$  with mean zero, it follows that  $if_{m-1, m-1}^{sym}(o_{i_1}, \dots, o_{i_{m-1}}; \theta)$  has influence func-

tion  $m \times if_{m, m}^{sym}(o_{i_1}, \dots, o_{i_{m-1}}, O; \theta) - \sum_{j=1}^{m-1} if_{m-1, m-1}^{sym}(o_{i_1}, \dots, o_{i_{j-1}}, O, o_{i_{j+1}}, \dots, o_{i_{m-1}}; \theta)$  on

the set  $\mathcal{O}_{m-1}$ . Thus  $m \times if_{m, m}^{sym}(o_{i_1}, \dots, o_{i_{m-1}}, O_{i_m}; \theta) = d_{m, \theta} \left[ if_{1, if_{m-1, m-1}^{sym}(o_{i_1}, \dots, o_{i_{m-1}}; \cdot)}(O_{i_m}; \theta) \right]$ .

Below  $f(O; \tilde{\theta}(\zeta))$ ,  $\zeta^T = (\zeta_1, \dots, \zeta_s)$  denotes an arbitrary smooth  $s$ -dimensional parametric submodel and  $\zeta_t$  denotes an arbitrary component of  $\zeta$ . ■

**Corollary 51** For  $m \geq 2$ ,

$$\Pi_\theta [\mathbb{IF}_{m-1,m-1,\setminus l_t}(\theta) |\mathcal{U}_{m-2}^\perp(\theta)] = -\Pi_\theta [\mathbb{IF}_{m,m,\setminus l_t}(\theta) |\mathcal{U}_{m-1}(\theta)] \quad (62)$$

$$\mathbb{IF}_{m,\setminus l_t}(\theta) = \Pi_\theta [\mathbb{IF}_{m,m,\setminus l_t}(\theta) |\mathcal{U}_{m-1}^\perp(\theta)] \quad (63)$$

$$\begin{aligned} & E_\theta [\mathbb{IF}_{m,\setminus l_{m+1}}(\theta) \tilde{\mathbb{S}}_{m,\bar{l}_m}(\theta)] \\ &= m! E_\theta \left( if_{m,m,\setminus l_{m+1}}^{sym}(O_{i_1}, \dots, O_{i_m}; \theta) \prod_{r=1}^m S_{l_r}(O_{i_r}, \theta) \right) \end{aligned} \quad (64)$$

**Proof.** (62): By lemma 50 and Theorem 5c(ii),

$$\begin{aligned} & \Pi_\theta [\mathbb{IF}_{m,m,\setminus l_t}(\theta) |\mathcal{U}_{m-1}(\theta)] \\ &= \mathbb{V} \left[ m E_\theta \left( IF_{m,m,\bar{l}_m}^{sym}(\theta) s_{l_t}(O_{i_m}) |O_{i_1}, \dots, O_{i_{m-1}} \right) \right] \\ &= \mathbb{V} \left[ m E_\theta \left( m^{-1} d_{m,\theta} \left\{ if_{1,if_{m-1,m-1}^{sym}}(O_{i_1}, \dots, O_{i_{m-1}}; \cdot) (O_{i_m}; \theta) \right\} s_{l_t}(O_{i_m}) |O_{i_1}, \dots, O_{i_{m-1}} \right) \right] \end{aligned}$$

Now, by part (iii) of lemma 50 and Eq. (4), the RHS is

$$\begin{aligned} & \mathbb{V} \left[ E_\theta \left( if_{1,if_{m-1,m-1}^{sym}}(O_{i_1}, \dots, O_{i_{m-1}}; \cdot) (O_{i_m}; \theta) s_{l_t}(O_{i_m}) |O_{i_1}, \dots, O_{i_{m-1}} \right) \right] \\ & - \mathbb{V} \left\{ E \left[ E \left[ (m-1) E \left[ if_{1,if_{m-1,m-1}^{sym}}(O_{i_1}, \dots, O_{i_{m-1}}; \cdot) (O_{i_m}; \theta) |O_{i_m}, O_{i_1}, \dots, O_{i_{m-2}} \right] s_{l_t}(O_{i_m}) |O_{i_1}, \dots, O_{i_{m-1}} \right] \right] \right\} \\ &= \mathbb{V} \left[ IF_{m-1,m-1,\setminus l_t}^{sym}(\theta) \right] - \mathbb{V} \left\{ (m-1) E_\theta \left[ IF_{m-1,m-1,\setminus l_t}^{sym}(\theta) |O_{i_1}, \dots, O_{i_{m-2}} \right] \right\} \end{aligned}$$

On the other hand, by part (iv) of the lemma 50,  $\Pi_\theta [\mathbb{IF}_{m-1,m-1,\setminus l_t}(\theta) |\mathcal{U}_{m-2}^\perp(\theta)] =$

$$\mathbb{V} [IF_{m-1,m-1,\setminus l_t}^{sym}(\theta)] - \mathbb{V} \left[ (m-1) E_\theta \left[ IF_{m-1,m-1,\setminus l_t,\bar{l}_{m-1}}^{sym}(\theta) |O_{i_1}, \dots, O_{i_{m-2}} \right] \right].$$

(63) : Write

$$\begin{aligned} \mathbb{IF}_{m,\setminus l_t}(\theta) &= \Pi_\theta [\mathbb{IF}_{m,m,\setminus l_t}(\theta) |\mathcal{U}_{m-1}^\perp(\theta)] + \{ \Pi [\mathbb{IF}_{2,2,\setminus l_t}(\theta) |\mathcal{U}_1(\theta)] + \mathbb{IF}_{1,\setminus l_t}(\theta) \} \\ &\quad + \sum_{j=2}^{m-1} \{ \Pi [\mathbb{IF}_{j+1,j+1,\setminus l_t}(\theta) |\mathcal{U}_j(\theta)] + \Pi [\mathbb{IF}_{jj,\setminus l_t}(\theta) |\mathcal{U}_{j-1}^\perp(\theta)] \} \end{aligned}$$

The RHS is  $\Pi_\theta [\mathbb{IF}_{m,m,\setminus l_t}(\theta) |\mathcal{U}_{m-1}^\perp(\theta)]$  by eq. (62).

$$(64) : E_\theta [\mathbb{IF}_{m,\setminus l_{m+1}}(\theta) \tilde{\mathbb{S}}_{m,\bar{l}_m}(\theta)] = E_\theta [\Pi [\mathbb{IF}_{m,m,\setminus l_{m+1}}(\theta) |\mathcal{U}_{m-1}^\perp(\theta)] \tilde{\mathbb{S}}_{m,\bar{l}_m}(\theta)] \text{ by}$$

eq. (63). But the RHS of this equation is the RHS of eq. (64). ■

**Proof. (Theorem 5c(i)):** By assumption  $\psi_{\setminus \bar{l}_{m-1}}(\theta) = E_\theta (\mathbb{IF}_{m-1}(\theta) \tilde{\mathbb{S}}_{m-1,\bar{l}_{m-1}}(\theta))$ .

Hence

$$\psi_{\setminus \bar{l}_m}(\theta) = E_\theta (\mathbb{IF}_{m-1}(\theta) \tilde{\mathbb{S}}_{m,\bar{l}_m}(\theta)) + E_\theta [\mathbb{IF}_{m-1,\setminus l_m}(\theta) \tilde{\mathbb{S}}_{m-1,\bar{l}_{m-1}}(\theta)]$$

By eq. (64), and the assumption  $if_{m-1,m-1}^{sym}(O_{i_1}, \dots, O_{i_m}; \theta)$  has an influence function,

we obtain

$$\begin{aligned} &E_\theta [\mathbb{IF}_{m-1,\setminus l_m}(\theta) \tilde{\mathbb{S}}_{m-1,\bar{l}_{m-1}}(\theta)] \\ &= (m-1)! E_\theta \left( if_{1if_{m-1,m-1}^{sym}(O_{i_1}, \dots, O_{i_{m-1}}; \cdot)}(O_{i_m}, \theta) S_{l_m}(O_{i_m}, \theta) \prod_{r=1}^{m-1} S_{l_r}(O_{i_r}, \theta) \right). \end{aligned}$$

We conclude that  $\mathbb{IF}_{m,m}$  exists and equals  $\mathbb{V} \left[ m^{-1} d_{m,\theta} \left\{ if_{1if_{m-1,m-1}^{sym}(O_{i_1}, \dots, O_{i_{m-1}}; \cdot)}(O_{i_m}, \theta) \right\} \right]$ .

■

Below is an alternative proof of Theorem 5c (i) and (ii)

**Proof.** First we show that for any  $j$ -dimensional parametric submodel  $\tilde{\theta}(\zeta)$ ,

$$\frac{\partial (\mathbb{IF}_{j,\psi}(\theta) + \psi(\theta))}{\partial l_j} \in \mathcal{U}_{j-1}^{\perp_{j,\theta}}(\theta)$$

where  $\mathcal{U}_0(\theta) = \phi$ . From eq (61) we know that

$$E_\theta \left( \frac{\partial (\mathbb{IF}_{j,\psi}(\theta) + \psi(\theta))}{\partial l_j} \tilde{\mathbb{S}}_{r,\bar{l}_r}(\theta) \right) = E_\theta \left( \frac{\partial (\mathbb{IF}_{j,\psi}(\theta))}{\partial l_r} \tilde{\mathbb{S}}_{r,\bar{l}_r}(\theta) \right) = 0$$

for all  $1 \leq r < j$ . Since  $\mathcal{M}(\Theta)$  is locally nonparametric, i.e.,  $\mathcal{U}_{j-1}(\theta) = \Gamma_{j-1}(\theta)$ , and

$$E_\theta \left( \frac{\partial (\mathbb{IF}_{j,\psi}(\theta) + \psi(\theta))}{\partial l_j} \right) = -E_\theta \left( \mathbb{IF}_{j,\psi}(\theta) \tilde{\mathbb{S}}_{1,l_j}(\theta) \right) + \psi_{\setminus l_j} = 0$$

therefore we have

$$\frac{\partial (\mathbb{IF}_{j,\psi}(\theta) + \psi(\theta))}{\partial l_j} \in \mathcal{U}_{j-1}^{\perp_{j,\theta}}(\theta) \quad (65)$$

ii) If  $\mathbb{IF}_m$  exists, we have  $\mathbb{IF}_m = \mathbb{V}(IF_m)$  with  $IF_m$  symmetric, such that

$$\begin{aligned} \psi_{\setminus \bar{l}_m}(\theta) &= E(\mathbb{IF}_m \mathbb{S}_{m,\bar{l}_m}) \\ &= E_\theta \left( \mathbb{V}[IF_m^c(\theta)] \mathbb{D}_m^{(\mathbb{S}_{m,\bar{l}_m})} \right) \\ &\quad + E_\theta \left( \mathbb{V}(mE[if_m(O_{i_1}, \dots, O_{i_m}; \theta) | O_{i_1}, \dots, O_{i_{m-1}}]^c) \mathbb{D}_{m-1}^{(\mathbb{S}_{m,\bar{l}_m})} \right) \\ &\quad + E(\mathbb{IF}_{m-2} \mathbb{S}_{m,\bar{l}_m}) \end{aligned}$$

Therefore,  $if_{m-1,m-1}(O_{i_1}, \dots, O_{i_{m-1}}; \theta) = mE[if_m(O_{i_1}, \dots, O_{i_m}; \theta) | O_{i_1}, \dots, O_{i_{m-1}}]^c \text{ wp1.}$

So that the lhs has an influence function since:

$$\begin{aligned}
& \{E_\theta[if_m(O_{i_1}, \dots, O_{i_m}; \theta) | o_{i_1}, \dots, o_{i_{m-1}}]^c\}_{\setminus l_m} \\
&= \{E_\theta[if_m(O_{i_1}, \dots, O_{i_m}; \theta) | o_{i_1}, \dots, o_{i_{m-1}}]\}_{\setminus l_m} \\
&+ \sum_{t=0}^{m-2} (-1)^{m-1-t} \sum_{\substack{i_{r_1} \neq i_{r_2} \dots \neq i_{r_t} \\ i_{r_t} \subset i_{m-1}}} E_\theta(if_m(O_{i_1}, O_{i_2}, \dots, O_{i_m}; \theta) | o_{i_{r_1}}, o_{i_{r_2}}, \dots, o_{i_{r_t}})_{\setminus l_m} \\
&= \{E_\theta[if_m(O_{i_1}, \dots, O_{i_m}; \theta) S_{l_m}(O_{i_m}) | o_{i_1}, \dots, o_{i_{m-1}}]\} \\
&+ \sum_{t=0}^{m-2} \left[ \begin{array}{c} (-1)^{m-1-t} (m-t) \\ \sum_{\substack{i_{r_1} \neq i_{r_2} \dots \neq i_{r_t} \\ i_{r_t} \subset i_{m-1}}} E \left( \begin{array}{c} E[if_m(O_{i_1}, \dots, O_{i_m}; \theta) | o_{i_{r_1}}, o_{i_{r_2}}, \dots, o_{i_{r_t}}] S_{l_{t+1}}(O_{i_{t+1}}) \\ | o_{i_{r_1}}, o_{i_{r_2}}, \dots, o_{i_{r_t}} \end{array} \right) \end{array} \right]
\end{aligned}$$

by equation 65.

i) if the first order influence function of  $if_{m-1,m-1,\theta}(O_{i_1}, \dots, O_{i_{m-1}})$ ,

$$if_{1,if_{m-1,m-1,\theta}(O_{i_1}, \dots, O_{i_{m-1}}; \cdot)}(O_{i_m}; \theta)$$

exists, then

$$\begin{aligned}
\psi_{\setminus \bar{l}_m}(\theta) &= E(\mathbb{IF}_{m-1} \mathbb{S}_{m, \bar{l}_m}) + \\
&E_\theta \left( (m-1)! if_{1,if_{m-1,m-1,\theta}(O_{i_1}, \dots, O_{i_{m-1}}; \cdot)}(O_{i_m}, \theta) \prod_{r=1}^m S_{l_r}(O_{i_r}, \theta) \right) \\
&= E_\theta \left( (m-1)! if_{1,if_{m-1,m-1,\theta}(O_{i_1}, \dots, O_{i_{m-1}}; \cdot)}^c(O_{i_m}, \theta) \prod_{r=1}^m S_{l_r}(O_{i_r}, \theta) \right)
\end{aligned}$$

If we switch the order of differentiating  $l_m$  with  $l_j$  ( $1 \leq j \leq m-1$ ), since

$$if_{m-1,m-1}(O_{i_1}, \dots, O_{i_{m-1}}, \theta)$$

is symmetric, we will have

$$\psi_{\setminus \bar{l}_m}(\theta) = E\left(\mathbb{IF}_{m-1}\mathbb{S}_{m,\bar{l}_m}\right) + E_\theta\left((m-1)!if_{1,if_{m-1,m-1,\theta}(O_{i_1},\dots,O_{i_m},O_{i_{m-1}};\cdot)}^c(O_{i_j},\theta)\prod_{r=1}^m S_{l_r}(O_{i_r},\theta)\right)$$

which means

$$E_\theta\left(\left[\begin{array}{c} if_{1,if_{m-1,m-1,\theta}(O_{i_1},\dots,O_{i_m},O_{i_{m-1}};\cdot)}^c(O_{i_m},\theta) \\ -if_{1,if_{m-1,m-1,\theta}(O_{i_1},\dots,O_{i_m},O_{i_{m-1}};\cdot)}^c(O_{i_j},\theta) \end{array}\right]\prod_{r=1}^m S_{l_r}(O_{i_r},\theta)\right) = 0$$

$$\iff if_{1,if_{m-1,m-1,\theta}(O_{i_1},\dots,O_{i_m},O_{i_{m-1}};\cdot)}^c(O_{i_m},\theta) = if_{1,if_{m-1,m-1,\theta}(O_{i_1},\dots,O_{i_m},O_{i_{m-1}};\cdot)}^c(O_{i_j},\theta) \text{ w.p.1}$$

therefore

$$\psi_{\setminus \bar{l}_m}(\theta) = E\left(\mathbb{IF}_{m-1}\mathbb{S}_{m,\bar{l}_m}\right) + E\left(\mathbb{IF}_{m,m}(\theta)\mathbb{S}_{m,\bar{l}_m}\right)$$

with

$$\mathbb{IF}_{m,m} \equiv \frac{1}{m}\mathbb{V}\left(if_{1,if_{m-1,m-1}(O_{i_1},\dots,O_{i_m},O_{i_{m-1}};\cdot)}^c(O_{i_m};\theta)\right)$$

■

**Proof.** (Proof of eq. (20)) We have proved in the text the following results that will

be used repeatedly throughout the proof:

$$v(x;\theta)K_{f_X,\infty}(x,X) = v(X;\theta)K_{f_X,\infty}(x,X),$$

$$\begin{aligned}
& if_{1,b(x_{i_1};\cdot)}(O_{i_2};\theta) \\
&= -E_\theta[H_1|x_{i_1}]^{-\frac{1}{2}} K_{f_X,\infty}(x_{i_1}, X_{i_2}) E_\theta[H_1|X_{i_2}]^{-\frac{1}{2}} \varepsilon_{b,i_2}(\theta) \\
&= -\frac{K_{Leb,\infty}(x_{i_1}, X_{i_2})}{g(x_{i_1})^{\frac{1}{2}} g(X_{i_2})^{\frac{1}{2}}} \varepsilon_{b,i_2}(\theta),
\end{aligned}$$

$$\begin{aligned}
& if_{1,p(x_{i_1};\cdot)}(O_{i_2};\theta) \\
&= -E_\theta[H_1|x_{i_1}]^{-\frac{1}{2}} K_{f_X,\infty}(x_{i_1}, X_{i_2}) E_\theta[H_1|X_{i_2}]^{-\frac{1}{2}} \varepsilon_{p,i_2}(\theta) \\
&= -\frac{K_{Leb,\infty}(x_{i_1}, X_{i_2})}{g(x_{i_1})^{\frac{1}{2}} g(X_{i_2})^{\frac{1}{2}}} \varepsilon_{p,i_2}(\theta),
\end{aligned}$$

and

$$IF_{1,K_{f_X,\infty}(X_{i_1},X_{i_2})} = -\{K_{f_X,\infty}(X_{i_1}, X_{i_3}) K_{f_X,\infty}(X_{i_3}, X_{i_2}) - K_{f_X,\infty}(X_{i_1}, X_{i_2})\}.$$

In addition, by an analogous argument, we can also show that

$$\begin{aligned}
& if_{1,g(x;\cdot)}(O) \\
&= IF_{1,E_\theta[H_1|x]} f_X(x) + E_\theta[H_1|x] IF_{1,f_X(x;\cdot)} \\
&= (H_1 - E[H_1|X]) K_{Leb,\infty}(X, x) + E_\theta[H_1|x] (K_{Leb,\infty}(X, x) - f_X(x)) \\
&= H_1 K_{Leb,\infty}(X, x) - g(x)
\end{aligned}$$

Now, we are ready to prove that eq. (20) holds for any  $m \geq 2$  by induction.

i) The case where  $m = 2$  was proved in the text.



ii) We now assume eq. (20) holds for  $m$  for some  $m \geq 2$ , and shall prove it is also true for  $m + 1$ . By assumption,

$$\begin{aligned}
& IF_{m,m,\psi,\bar{i}_m}(\theta) \\
&= \varepsilon_{b,i_1}(\theta) g(X_{i_1})^{-\frac{1}{2}} \left[ \begin{aligned} & \sum_{j=0}^{m-2} c(m,j) \times \\ & \prod_{s=1}^j \frac{H_{1,i_{s+1}}}{g(X_{i_{s+1}})} K_{Leb,\infty}(X_{i_s}, X_{i_{s+1}}) \\ & \times K_{Leb,\infty}(X_{i_{j+1}}, X_{i_m}) \end{aligned} \right] g(X_{i_m})^{-\frac{1}{2}} \varepsilon_{p,i_m}(\theta)
\end{aligned}$$

Following the results from part 5c) of Theorem 3,  $IF_{m+1,m+1,\psi,\bar{i}_m}(\theta)$  exists if and only if  $if_{1,IF_{m,m,\psi,\bar{i}_m}}(\cdot)$  exists. By the chain rule,

$$\begin{aligned}
& if_{1,if_{m,m,\psi,\bar{i}_m}}(\mathbf{O}_{i_m} \cdot) (O_{i_{m+1}}; \theta) \\
&= \left\{ \begin{aligned} & \left( \begin{aligned} & H_{1,i_1} if_{1,\varepsilon_{b,i_1}}(\cdot) (O_{i_{m+1}}) \varepsilon_{p,i_m}(\theta) + \\ & H_{1,i_m} \varepsilon_{b,i_1}(\theta) if_{1,\varepsilon_{p,i_m}}(\cdot) (O_{i_{m+1}}) \end{aligned} \right) g(X_{i_1})^{-\frac{1}{2}} g(X_{i_m})^{-\frac{1}{2}} \\ & - \frac{1}{2} \varepsilon_{b,i_1}(\theta) \varepsilon_{p,i_m}(\theta) g(X_{i_1})^{-\frac{1}{2}} g(X_{i_m})^{-\frac{1}{2}} \times \\ & \left[ \frac{if_{1,g(X_{i_1})}(O_{i_{m+1}})}{g(X_{i_1})} + \frac{if_{1,g(X_{i_m})}(O_{i_{m+1}})}{g(X_{i_m})} \right] \end{aligned} \right\} \times \\
& \left[ \begin{aligned} & \sum_{j=0}^{m-2} c(m,j) \times \\ & \prod_{s=1}^j \frac{H_{1,i_{s+1}}}{g(X_{i_{s+1}})} K_{Leb,\infty}(X_{i_s}, X_{i_{s+1}}) \\ & \times K_{Leb,\infty}(X_{i_{j+1}}, X_{i_m}) \end{aligned} \right] \\
& - \varepsilon_{b,i_1}(\theta) \varepsilon_{p,i_m}(\theta) g(X_{i_1})^{-\frac{1}{2}} g(X_{i_m})^{-\frac{1}{2}} \sum_{j=0}^{m-2} c(m,j) \times \\
& \left[ \begin{aligned} & \sum_{t=1}^j \frac{H_{1,i_{t+1}} K_{Leb,\infty}(X_{i_t}, X_{i_{t+1}})}{g^2(X_{i_{t+1}})} if_{1,g(X_{i_{t+1}})}(O_{i_{m+1}}) \times \\ & \prod_{s \neq t} \frac{H_{1,i_{s+1}}}{g(X_{i_{s+1}})} K_{Leb,\infty}(X_{i_s}, X_{i_{s+1}}) K_{Leb,\infty}(X_{i_{j+1}}, X_{i_m}) \end{aligned} \right] \\
&=
\end{aligned}$$

$$\begin{aligned}
& \left\{ \begin{aligned} & -\frac{H_{1,i_1} K_{Leb,\infty}(X_{i_1}, X_{i_{m+1}})}{g(X_{i_1})g(X_{i_{m+1}})^{\frac{1}{2}}g(X_{i_m})^{\frac{1}{2}}} \varepsilon_{b,i_{m+1}}(\theta) \varepsilon_{p,i_m}(\theta) \\ & -\varepsilon_{b,i_1}(\theta) \frac{H_{1,i_m} K_{Leb,\infty}(X_{i_m}, X_{i_{m+1}})}{g(X_{i_m})g(X_{i_1})^{\frac{1}{2}}g(X_{i_{m+1}})^{\frac{1}{2}}} \varepsilon_{p,i_{m+1}} \\ & -\frac{1}{2} \varepsilon_{b,i_1}(\theta) \varepsilon_{p,i_m}(\theta) g(X_{i_1})^{-\frac{1}{2}} g(X_{i_m})^{-\frac{1}{2}} \times \\ & \left[ \frac{H_{1,i_{m+1}} K_{Leb,\infty}(X_{i_{m+1}}, X_{i_1})}{g(X_{i_1})} + \frac{H_{1,i_{m+1}} K_{Leb,\infty}(X_{i_{m+1}}, X_{i_m})}{g(X_{i_m})} - 2 \right] \end{aligned} \right\} \\
& \times \left[ \begin{aligned} & \sum_{j=0}^{m-2} c(m, j) \times \\ & \prod_{s=1}^j \frac{H_{1,i_{s+1}}}{g(X_{i_{s+1}})} K_{Leb,\infty}(X_{i_s}, X_{i_{s+1}}) \\ & \times K_{Leb,\infty}(X_{i_{j+1}}, X_{i_m}) \end{aligned} \right] \\
& - \varepsilon_{b,i_1}(\theta) \varepsilon_{p,i_m}(\theta) g(X_{i_1})^{-\frac{1}{2}} g(X_{i_m})^{-\frac{1}{2}} \sum_{j=0}^{m-2} c(m, j) \times \\
& \left[ \begin{aligned} & \sum_{t=1}^j \frac{H_{1,i_{t+1}} K_{Leb,\infty}(X_{i_t}, X_{i_{t+1}})}{g^2(X_{i_{t+1}})} (H_{1,i_{m+1}} K_{Leb,\infty}(X_{i_{m+1}}, X_{i_{t+1}}) - g(X_{i_{t+1}})) \times \\ & \prod_{s \neq t} \frac{H_{1,i_{s+1}}}{g(X_{i_{s+1}})} K_{Leb,\infty}(X_{i_s}, X_{i_{s+1}}) K_{Leb,\infty}(X_{i_{j+1}}, X_{i_m}) \end{aligned} \right] \\
& =
\end{aligned}$$

$$\begin{aligned}
& + \sum_{j=0}^{m-2} (-1)^j \binom{m-2}{j} \left( \begin{aligned} & \varepsilon_{b,i_{m+1}}(\theta) g(X_{i_{m+1}})^{-\frac{1}{2}} \frac{H_{1,i_1} K_{Leb,\infty}(X_{i_{m+1}}, X_{i_1})}{g(X_{i_1})} \\ & \prod_{s=1}^j \frac{H_{1,i_{s+1}}}{g(X_{i_{s+1}})} K_{Leb,\infty}(X_{i_s}, X_{i_{s+1}}) K_{Leb,\infty}(X_{i_{j+1}}, X_{i_m}) \\ & \times g(X_{i_m})^{-\frac{1}{2}} \varepsilon_{p,i_m}(\theta) \end{aligned} \right) \\
& + \sum_{j=0}^{m-2} (-1)^j \binom{m-2}{j} \left( \begin{aligned} & \varepsilon_{b,i_1}(\theta) g(X_{i_1})^{-\frac{1}{2}} \prod_{s=1}^j \frac{H_{1,i_{s+1}}}{g(X_{i_{s+1}})} K_{Leb,\infty}(X_{i_s}, X_{i_{s+1}}) \\ & \frac{H_{1,i_m} K_{Leb,\infty}(X_{i_{j+1}}, X_{i_m})}{g(X_{i_m})} K_{Leb,\infty}(X_{i_m}, X_{i_{m+1}})^{\frac{1}{2}} \\ & \times g(X_{i_{m+1}})^{-\frac{1}{2}} \varepsilon_{p,i_{m+1}} \end{aligned} \right) \\
& + \frac{1}{2} \sum_{j=0}^{m-2} (-1)^j \binom{m-2}{j} \left\{ \begin{aligned} & \varepsilon_{b,i_1}(\theta) \varepsilon_{p,i_m}(\theta) \left( \begin{aligned} & \frac{H_{1,i_{m+1}} K_{Leb,\infty}(X_{i_{m+1}}, X_{i_1})}{g(X_{i_1})} \\ & + \frac{H_{1,i_{m+1}} K_{Leb,\infty}(X_{i_{m+1}}, X_{i_{j+1}})}{g(X_{i_{m+1}})} \end{aligned} \right) \\ & \times \prod_{s=1}^j \frac{H_{1,i_{s+1}}}{g(X_{i_{s+1}})} K_{Leb,\infty}(X_{i_s}, X_{i_{s+1}}) \times \\ & K_{Leb,\infty}(X_{i_{j+1}}, X_{i_m}) g(X_{i_1})^{-\frac{1}{2}} g(X_{i_m})^{-\frac{1}{2}} \end{aligned} \right\} \\
& + IF_{m,m,\psi,\bar{i}_m}(\theta) \\
& + \sum_{j=0}^{m-2} (-1)^j \binom{m-2}{j} \left[ \sum_{t=1}^j \left( \begin{aligned} & \varepsilon_{b,i_1}(\theta) \varepsilon_{p,i_m}(\theta) g(X_{i_1})^{-\frac{1}{2}} g(X_{i_m})^{-\frac{1}{2}} \times \\ & \frac{H_{1,i_{t+1}} K_{Leb,\infty}(X_{i_t}, X_{i_{t+1}})}{g(X_{i_{t+1}})} H_{1,i_{m+1}} \frac{K_{Leb,\infty}(X_{i_{t+1}}, X_{i_{m+1}})}{g(X_{i_{m+1}})} \\ & \times \prod_{s \neq t} \frac{H_{1,i_{s+1}}}{g(X_{i_{s+1}})} K_{Leb,\infty}(X_{i_s}, X_{i_{s+1}}) K_{Leb,\infty}(X_{i_{j+1}}, X_{i_m}) \end{aligned} \right) \right] \\
& - \sum_{j=0}^{m-2} (-1)^j \binom{m-2}{j} j \left( \begin{aligned} & \varepsilon_{b,i_1}(\theta) \varepsilon_{p,i_m}(\theta) g(X_{i_1})^{-\frac{1}{2}} g(X_{i_m})^{-\frac{1}{2}} \times \\ & \prod_{s=1}^j \frac{H_{1,i_{s+1}}}{g(X_{i_{s+1}})} K_{Leb,\infty}(X_{i_s}, X_{i_{s+1}}) K_{Leb,\infty}(X_{i_{j+1}}, X_{i_m}) \end{aligned} \right)
\end{aligned}$$

Applying the operator  $d_{m+1,\theta}$ , which is defined in Eq. (4), on the statistic above,

it is straightforward to show that

$$\begin{aligned}
\mathbb{IF}_{m+1,m+1,\psi,\bar{i}_{m+1}}(\theta) &= \frac{1}{m+1} \left( \mathbb{V} \left\{ d_{m+1,\theta} \left[ if_{1,IF_{m,m,\psi,\bar{i}_m}}(\cdot)(\theta) \right] \right\} \right) \\
&= \mathbb{V} \left\{ d_{m+1,\theta} \left[ \begin{aligned} &\varepsilon_{b,i_1}(\theta) g(X_{i_1})^{-\frac{1}{2}} \prod_{s=1}^{m-1} \frac{H_{1,i_{s+1}}}{g(X_{i_{s+1}})} K_{Leb,\infty}(X_{i_s}, X_{i_{s+1}}) \\ &\times K_{Leb,\infty}(X_{i_m}, X_{i_{m+1}}) g(X_{i_{m+1}})^{-\frac{1}{2}} \varepsilon_{p,i_{m+1}}(\theta) \end{aligned} \right] \right\} \\
&= \mathbb{V} \left\{ \begin{aligned} &\varepsilon_{b,i_1}(\theta) g(X_{i_1})^{-\frac{1}{2}} \left[ \begin{aligned} &\sum_{j=0}^{m-1} c(m+1, j) \times \\ &\prod_{s=1}^j \frac{H_{1,i_{s+1}}}{g(X_{i_{s+1}})} K_{Leb,\infty}(X_{i_s}, X_{i_{s+1}}) \\ &\times K_{Leb,\infty}(X_{i_{j+1}}, X_{i_{m+1}}) \end{aligned} \right] \\ &\times g(X_{i_{m+1}})^{-\frac{1}{2}} \varepsilon_{p,i_{m+1}}(\theta) \end{aligned} \right\}
\end{aligned}$$

■

**Proof.** (Theorem 15) By eq (28) and eq (29) and part a of the preceding lemma

$$\begin{aligned}
\tilde{\eta}_k &= -E \left[ \dot{B} \dot{P} H_1 \bar{Z}_k \bar{Z}_k^T \right]^{-1} E \left[ \bar{Z}_k \dot{P} \left( \hat{B} - B \right) H_1 \right] \\
\tilde{\alpha}_k &= -E \left[ \dot{B} \dot{P} H_1 \bar{Z}_k \bar{Z}_k^T \right]^{-1} E \left[ \bar{Z}_k \dot{B} \left( \hat{P} - P \right) H_1 \right]
\end{aligned}$$

and hence

$$\begin{aligned}
\tilde{B} - \hat{B} &= -\dot{B} \bar{Z}_k^T E \left[ \dot{B} \dot{P} H_1 \bar{Z}_k \bar{Z}_k^T \right]^{-1} E \left[ \dot{P} \dot{B} \bar{Z}_k \left( B - \hat{B} \right) \left\{ \dot{B} \right\}^{-1} H_1 \right] \\
\tilde{P} - \hat{P} &= -\dot{P} \bar{Z}_k^T E \left[ \dot{B} \dot{P} H_1 \bar{Z}_k \bar{Z}_k^T \right]^{-1} E \left[ \dot{B} \dot{P} \bar{Z}_k \left( P - \hat{P} \right) \left\{ \dot{P} \right\}^{-1} H_1 \right]
\end{aligned}$$

Thus

$$\begin{aligned}
Q \frac{\tilde{P} - \hat{P}}{\dot{P}} &= -Q \bar{Z}_k^T E_\theta \left[ Q^2 \bar{Z}_k \bar{Z}_k^T \right]^{-1} E \left[ Q^2 \bar{Z}_k \frac{(P - \hat{P})}{\dot{P}} \right] \\
&= \Pi \left[ \frac{P - \hat{P}}{\dot{P}} Q | Q \bar{Z}_k \right] \\
Q \frac{\tilde{B} - \hat{B}}{\dot{B}} &= -Q \bar{Z}_k^T E_\theta \left[ Q^2 \bar{Z}_k \bar{Z}_k^T \right]^{-1} E \left[ Q^2 \bar{Z}_k \frac{(B - \hat{B})}{\dot{B}} \right] \\
&= \Pi \left[ \left( \frac{B - \hat{B}}{\dot{B}} \right) Q | Q \bar{Z}_k \right]
\end{aligned}$$

and hence

$$\begin{aligned}
Q \left( \frac{(P - \tilde{P})}{\dot{P}} \right) &= \Pi^\perp \left[ \left( \frac{P - \hat{P}}{\dot{P}} \right) Q | Q \bar{Z}_k \right] \\
Q \left( \frac{(B - \tilde{B})}{\dot{B}} \right) &= \Pi^\perp \left[ \left( \frac{B - \hat{B}}{\dot{B}} \right) Q | Q \bar{Z}_k \right]
\end{aligned}$$

But by the previous theorem,

$$TB_k = E \left[ \left\{ \tilde{B} - B \right\} \left\{ \tilde{P} - P \right\} H_1 \right] = E \left[ Q \left( \frac{B - \tilde{B}}{\dot{B}} \right) Q \left( \frac{(P - \tilde{P})}{\dot{P}} \right) \right]$$

proving the theorem. ■

**Proof.** (Theorem 18) Under our assumptions, the following holds uniformly for

$\theta \in \Theta$ .

$$\begin{aligned}
TB_k^2 &= \left\{ E \left[ \Pi^\perp \left[ \left( \frac{P - \hat{P}}{\dot{P}} \right) Q | Q \bar{Z}_k \right] \Pi^\perp \left[ \left( \frac{B - \hat{B}}{\dot{B}} \right) Q | Q \bar{Z}_k \right] \right] \right\}^2 \\
&\leq E \left\{ \Pi^\perp \left[ \left( \frac{P - \hat{P}}{\dot{P}} \right) Q | Q \bar{Z}_k \right]^2 \right\} \times E \left\{ \Pi^\perp \left[ \left( \frac{B - \hat{B}}{\dot{B}} \right) Q | Q \bar{Z}_k \right]^2 \right\}
\end{aligned}$$

by Cauchy Schwartz. Now

$$\begin{aligned}
& E \left\{ \Pi^\perp \left[ \left( \frac{P - \hat{P}}{\dot{P}} \right) Q | Q \bar{Z}_k \right]^2 \right\} \\
&= \inf_{\varsigma_l} \int_{R^d} Q^2 \left( \frac{p(X) - \hat{p}(X)}{\dot{p}(X)} - \sum_{l=1}^k \varsigma_l z_l(X) \right)^2 f(X) dX \\
&= \inf_{\varsigma_l} \int_{R^d} Q^2 \left( \frac{(p(X) - \hat{p}(X))}{\dot{p}(X)} - \sum_{l=1}^k \varsigma_l^* \varphi_l(X) \right)^2 Q^2 f(X) dX \\
&= \inf_{\varsigma_l} \int_{R^d} \left( (p(X) - \hat{p}(X)) - \sum_{l=1}^k \varsigma_l^* \varphi_l(X) \right)^2 Q^2 f(X) dX \\
&\leq \|Q^2 f(X)\|_\infty \inf_{\varsigma_l} \int_{R^d} \left( (p(X) - \hat{p}(X)) - \sum_{l=1}^k \varsigma_l^* \varphi_l(X) \right)^2 dX \\
&\leq \|Q^2 f(X)\|_\infty O_p(k^{-2\beta_p/d}) = O_p(k^{-2\beta_p/d})
\end{aligned}$$

The last equality follows from the fact that under the stated assumptions  $\|Q^2 f(X)\|_\infty = O(1)$ . Similarly  $E \left\{ \Pi^\perp \left[ \left( \frac{B - \hat{B}}{\dot{B}} \right) Q | Q \bar{Z}_k \right]^2 \right\} = O_p(k^{-2\beta_b/d})$ .

■

**Proof.** (Theorem 20) By theorem 19

$$\begin{aligned}
IF_{1, \tilde{\psi}_k, i_i}(\theta) &= if_{1, \tilde{\psi}_k}(O_{i_1}, \theta) \\
&= H_{i_i}(\tilde{b}(\theta), \tilde{p}(\theta)) - \tilde{\psi}_k(\theta) \\
&= h(O_{i_1}, \tilde{b}(X_{i_1}, \theta), \tilde{p}(X_{i_1}, \theta)) - \tilde{\psi}_k(\theta)
\end{aligned}$$

and by part 5.c of theorem 3

$$\mathbb{V} \left[ IF_{22, \tilde{\psi}_k, \bar{i}_2} \right] = \frac{1}{2} \left\{ \Pi_\theta \left[ \mathbb{V} \left[ IF_{1, if_{1, \tilde{\psi}_k}(O_{i_1, \cdot}), i_2}(\theta) \right] | \mathcal{U}_1^{\perp, 2}(\theta) \right] \right\}.$$

Now

$$\begin{aligned} IF_{1,if_{1,\tilde{\psi}_k},(O_{i_1},\theta),i_2}(\theta) &= h_{\tilde{b}}\left(O_{i_1},\tilde{b}(X_{i_1},\theta),\tilde{p}(X_{i_1},\theta)\right) IF_{1,\tilde{b}(X_{i_1},\cdot),i_2}(\theta) \\ &\quad + h_{\tilde{p}}\left(O_{i_1},\tilde{b}(X_{i_1},\theta),\tilde{p}(X_{i_1},\theta)\right) IF_{1,\tilde{p}(X_{i_1},\cdot),i_2}(\theta) \end{aligned}$$

where

$$\begin{aligned} h_{\tilde{b}}\left(O_{i_1},\tilde{b}(X_{i_1},\theta),\tilde{p}(X_{i_1},\theta)\right) &= H_{1i_1}\tilde{p}(X_{i_1},\theta) + H_{2i_1} \\ h_{\tilde{p}}\left(O_{i_1},\tilde{b}(X_{i_1},\theta),\tilde{p}(X_{i_1},\theta)\right) &= H_{1i_1}\tilde{b}(X_{i_1},\theta) + H_{3i_1} \end{aligned}$$

$$\begin{aligned} IF_{1,\tilde{b}(X_{i_1},\cdot),i_2}(\theta) &= IF_{1,b^*(X_{i_1},\tilde{\eta}_k(\cdot)),i_2}(\theta) \\ &= \dot{B}_{i_1}\overline{Z}_{ki_1}^T IF_{1,\tilde{\eta}_k(\cdot),i_2}(\theta) \\ &= -\dot{B}\overline{Z}_{ki_1}^T \left\{E_\theta \left[\dot{P}\dot{B}H_1\overline{Z}_k\overline{Z}_k^T\right]\right\}^{-1} \left[\left\{H_1\tilde{b}(X,\theta) + H_3\right\}\dot{P}\overline{Z}_k\right]_{i_2}, \end{aligned}$$

and

$$IF_{1,\tilde{p}(X_{i_1},\cdot),i_2}(\theta) = -\dot{P}_{i_1}\overline{Z}_{ki_1}^T \left\{E_\theta \left[\dot{P}\dot{B}H_1\overline{Z}_k\overline{Z}_k^T\right]\right\}^{-1} \left[\left\{H_1\tilde{p}(X,\theta) + H_2\right\}\dot{B}\overline{Z}_k\right]_{i_2}.$$

$$\begin{aligned} IF_{1,if_{1,\tilde{\psi}_k},(O_{i_1},\cdot),i_2}(\theta) &= -\left\{H_1\tilde{p}(X,\theta) + H_2\right\}_{i_1} \dot{B}_{i_1}\overline{Z}_{ki_1}^T \left\{E_\theta \left[\dot{P}\dot{B}H_1\overline{Z}_k\overline{Z}_k^T\right]\right\}^{-1} \\ &\quad \times \left[\left\{H_1\tilde{b}(X,\theta) + H_3\right\}\dot{P}\overline{Z}_k\right]_{i_2} \\ &\quad - \left\{H_1\tilde{b}(X,\theta) + H_3\right\}_{i_1} \dot{P}_{i_1}\overline{Z}_{ki_1}^T \left\{E_\theta \left[\dot{P}\dot{B}H_1\overline{Z}_k\overline{Z}_k^T\right]\right\}^{-1} \\ &\quad \times \left[\left\{H_1\tilde{p}(X,\theta) + H_2\right\}\dot{B}\overline{Z}_k\right]_{i_2} \end{aligned}$$



and further

$$\Pi_\theta \left[ \mathbb{V} \left[ IF_{1,if_{1,\tilde{\psi}_k},(O_{i_1},\cdot),i_2}(\theta) \mid \mathcal{U}_1(\theta) \right] \right] = 0$$

since

$$E_\theta \left[ \{H_1\tilde{p}(X, \theta) + H_2\} \dot{B}\bar{Z}_k \right] = E_\theta \left[ \{H_1\tilde{b}(X, \theta) + H_3\} \dot{P}\bar{Z}_k \right] = 0$$

and thus  $IF_{1,if_{1,\tilde{\psi}_k},(O_{i_1},\cdot),i_2}(\theta)$  is degenerate. Because  $IF_{1,if_{1,\tilde{\psi}_k},(O_{i_1},\cdot),i_2}(\theta)$  has two terms, it appears that  $IF_{22,\tilde{\psi}_k,\bar{i}_2}$  will consist of two terms. However by the symmetry upon interchange of  $i_2$  and  $i_1$ , and the permutation invariance of the operator  $\mathbb{V}$

$$\begin{aligned} & \mathbb{V} \left[ IF_{1,if_{1,\tilde{\psi}_k},(O_{i_1},\cdot),i_2}(\theta) \right] \\ &= \mathbb{V} \left[ \begin{aligned} & -2 \{H_1\tilde{p}(X, \theta) + H_2\}_{i_1} \dot{B}_{i_1} \bar{Z}_{ki_1}^T \left\{ E_\theta \left[ \dot{P}\dot{B}H_1\bar{Z}_k\bar{Z}_k^T \right] \right\}^{-1} \\ & \times \left[ \bar{Z}_k \left\{ H_1\tilde{b}(X, \theta) + H_3 \right\} \dot{P} \right]_{i_2} \end{aligned} \right] \end{aligned}$$

Thus we can take

$$\begin{aligned} IF_{22,\tilde{\psi}_k,\bar{i}_2} &= - \{H_1\tilde{p}(X, \theta) + H_2\}_{i_1} \dot{B}_{i_1} \bar{Z}_{ki_1}^T \left\{ E_\theta \left[ \dot{P}\dot{B}H_1\bar{Z}_k\bar{Z}_k^T \right] \right\}^{-1} \\ & \times \left[ \bar{Z}_k \left\{ H_1\tilde{b}(X, \theta) + H_3 \right\} \dot{P} \right]_{i_2} \end{aligned}$$

as was to be proved. We now complete the proof of the Theorem by induction. We

assume it is true for  $IF_{mm,\tilde{\psi}_k,\bar{i}_m}$  and prove it is true for  $IF_{(m+1)(m+1),\tilde{\psi}_k,\bar{i}_{m+1}}$ . Now

$$\mathbb{V} \left[ IF_{(m+1)(m+1),\tilde{\psi}_k,\bar{i}_{m+1}}(\theta) \right] = \frac{1}{m} \mathbb{V} \left[ \Pi_\theta \left[ IF_{1,if_{mm,\tilde{\psi}_k},(O_{\bar{i}_m},\cdot),i_{m+1}}(\theta) \mid \mathcal{U}_m^{\perp\theta,m+1}(\theta) \right] \right]$$

Now by the induction hypothesis,

$$\begin{aligned}
& if_{mm, \tilde{\psi}_k, (O_{\tilde{i}_m}, \theta) \\
&= (-1)^{m-1} \left[ \left( H_1 \tilde{P}(\theta) + H_2 \right) \dot{B} \bar{Z}_k^T \right]_{i_1} \\
&\times \left[ \prod_{s=3}^m \left\{ E_\theta \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} \left\{ \begin{array}{c} \left( \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right)_{i_s} \\ -E_\theta \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \end{array} \right\} \right] \\
&\times \left\{ E_\theta \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} \left[ \bar{Z}_k \left( H_1 \tilde{B}(\theta) + H_3 \right) \dot{P} \right]_{i_2}
\end{aligned}$$

The derivatives with respect to the  $\theta$ 's in  $\tilde{P}(\theta), \tilde{B}(\theta)$  and in the  $m-1$  terms  $\left\{ E_\theta \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1}$  will each contribute a term to  $\mathbb{V} \left[ IF_{(m+1)(m+1), \tilde{\psi}_k, \tilde{i}_{m+1}}(\theta) \right]$ . However differentiating wrt to the  $\theta$  in the  $m-2$  terms  $E_\theta \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right]$  will not contribute to  $\mathbb{V} \left[ IF_{(m+1)(m+1), \tilde{\psi}_k, \tilde{i}_{m+1}}(\theta) \right]$  as the contribution from each of these  $m-2$  terms to  $IF_{1, if_{mm, \tilde{\psi}_k, (O_{\tilde{i}_m}, \cdot), i_{m+1}}(\theta)}$  is only a function of  $m$  units' data and is thus an element of  $\mathcal{U}_m(\theta)$  which is orthogonal to the space  $\mathcal{U}_m^{\perp, m+1}(\theta)$  that is projected on..

Now

$$\begin{aligned}
& IF_{1, \left\{ E_\theta \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1}, i_{m+1}}(\theta) \\
&= - \left\{ E_\theta \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} \left\{ \begin{array}{c} \left( \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right)_{i_{m+1}} \\ -E_\theta \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \end{array} \right\} \left\{ E_\theta \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1}
\end{aligned}$$

so upon permuting the unit indices, the contribution of each of these  $m - 1$  terms to

$IF_{1,if_{mm,\tilde{\psi}_k},(O_{\tilde{i}_m},\cdot),i_{m+1}}(\theta)$  is

$$\begin{aligned}
& - (-1)^{m-1} \left[ \left( H_1 \tilde{P}(\theta) + H_2 \right) \dot{B} \bar{Z}_k^T \right]_{i_1} \\
& \times \left[ \prod_{s=3}^{m+1} \left\{ E_\theta \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} \left\{ \begin{array}{c} \left( \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right)_{i_s} \\ - E_\theta \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \end{array} \right\} \right] \\
& \times \left\{ E_\theta \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} \left[ \bar{Z}_k \left( H_1 \tilde{B}(\theta) + H_3 \right) \dot{P} \right]_{i_2}
\end{aligned} \tag{66}$$

which is already degenerate ( i.e., orthogonal to  $\mathcal{U}_m(\theta)$ ). Differentiating with respect to the  $\theta$ 's of  $\tilde{P}(\theta), \tilde{B}(\theta)$  in  $IF_{1,if_{mm,\tilde{\psi}_k},(O_{\tilde{i}_m},\cdot),i_{m+1}}(\theta)$  we obtain

$$\begin{aligned}
& = (-1)^{m-1} IF_{1,\tilde{b}(X_{i_1},\cdot),i_{m+1}}(\theta) \left[ H_1 \dot{B} \bar{Z}_k^T \right]_{i_1} \\
& \left[ \prod_{s=3}^m \left\{ E_\theta \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} \left\{ \left( \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right)_{i_s} - E_\theta \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\} \right] \times \\
& \left\{ E_\theta \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} \left[ \bar{Z}_k \left( H_1 \tilde{B}(\theta) + H_3 \right) \dot{P} \right]_{i_2} \\
& + (-1)^{m-1} \left[ \left( H_1 \tilde{P}(\theta) + H_2 \right) \dot{B} \bar{Z}_k^T \right]_{i_1} \times \\
& \left[ \prod_{s=3}^m \left\{ E_\theta \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} \left\{ \left( \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right)_{i_s} - E_\theta \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\} \right] \times \\
& \left[ \bar{Z}_k H_1 \dot{P} \right] IF_{1,\tilde{p}(X_{i_2},\cdot),i_{m+1}}(\theta)
\end{aligned}$$

Substituting in the above expressions for  $IF_{1,\tilde{b}(X_{i_1},\cdot),i_{m+1}}(\theta)$  and  $IF_{1,\tilde{p}(X_{i_2},\cdot),i_{m+1}}(\theta)$ , then projecting on  $\mathcal{U}_m^{\perp\theta,m+1}(\theta)$ , and again permuting unit indices, we obtain two identical terms both equal to eq (66) . Thus we obtain  $m + 1$  identical terms in

all. Upon dividing by  $m + 1$ , we conclude that  $\mathbb{V} \left[ IF_{(m+1)(m+1), \tilde{\psi}_k, \tilde{i}_{m+1}}((\theta)) \right]$  equals  $\mathbb{V}$  operating on (66), proving the theorem.

■

**Proof.** (Theorem 23) Equation (34) follows from eq. (33) by our assumption of rate optimality of the initial estimators. We next prove eq. (32) by induction. For  $m = 1$ .

$$\begin{aligned}
EB_1 &= E \left[ \hat{\psi}_{1,k} \right] - \tilde{\psi}_k \\
&= E \left[ \hat{B} \hat{P} H_1 + \hat{B} H_2 + \hat{P} H_3 \right] - E \left[ \tilde{B} \tilde{P} H_1 + \tilde{B} H_2 + \tilde{P} H_3 \right] \\
&= E \left[ \dot{B} \dot{P} \bar{Z}_k^T (P - \hat{P}) \left\{ \dot{P} \right\}^{-1} H_1 \right] E_\theta \left[ \dot{B} \dot{P} H_1 \bar{Z}_k \bar{Z}_k^T \right]^{-1} \\
&\quad \times E \left[ \dot{P} \dot{B} \bar{Z}_k (B - \hat{B}) \left\{ \dot{B} \right\}^{-1} H_1 \right]
\end{aligned}$$

where the last equality follows from

$$\begin{aligned}
\tilde{B} &= \hat{B} - B \bar{Z}_k^T E_\theta \left[ \dot{B} \dot{P} H_1 \bar{Z}_k \bar{Z}_k^T \right]^{-1} E \left[ \dot{P} \dot{B} \bar{Z}_k (B - \hat{B}) \left\{ \dot{B} \right\}^{-1} H_1 \right] \\
\tilde{P} &= \hat{P} - \dot{P} \bar{Z}_k^T E_\theta \left[ \dot{B} \dot{P} H_1 \bar{Z}_k \bar{Z}_k^T \right]^{-1} E \left[ \dot{B} \dot{P} \bar{Z}_k (P - \hat{P}) \left\{ \dot{P} \right\}^{-1} H_1 \right]
\end{aligned}$$

Next, we proceed by induction. Assume 32 holds for  $m - 1 \geq 1$ , we next show that it holds for  $m$ .

$$\begin{aligned}
EB_m &= EB_{m-1} + E \left[ IF_{mm, \tilde{\psi}_k, \tilde{i}_j} \right] \\
&= \left\{ \begin{aligned} &(-1)^{m-2} E \left[ Q^2 \left( \frac{B-\hat{B}}{\dot{B}} \right) \bar{Z}_k^T \right] \\ &\left\{ E \left[ Q^2 \bar{Z}_k \bar{Z}_k^T \right] - I_{k \times k} \right\}^{m-2} \left\{ E \left[ Q^2 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} E \left[ Q^2 \left( \frac{P-\hat{P}}{\dot{P}} \right) \right] \end{aligned} \right\} \\
&+ \left\{ \begin{aligned} &(-1)^{m-1} E \left[ Q^2 \left( \frac{B-\hat{B}}{\dot{B}} \right) \bar{Z}_k^T \right] \\ &\times \left\{ E \left[ Q^2 \bar{Z}_k \bar{Z}_k^T \right] - I_{k \times k} \right\}^{m-2} E \left[ Q^2 \left( \frac{P-\hat{P}}{\dot{P}} \right) \right] \end{aligned} \right\} \\
&= (-1)^{m-1} E \left[ Q^2 \left( \frac{B-\hat{B}}{\dot{B}} \right) \bar{Z}_k^T \right] \left\{ E \left[ Q^2 \bar{Z}_k \bar{Z}_k^T \right] - I_{k \times k} \right\}^{m-1} \\
&\times \left\{ E \left[ Q^2 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} E \left[ Q^2 \left( \frac{P-\hat{P}}{\dot{P}} \right) \right]
\end{aligned}$$

Finally we prove that 32 implies 33. For any random variable  $H$  define

$$\hat{R}(H) = \hat{\Pi} \left[ \delta g H | \hat{Q} \bar{Z}_k \right] = \hat{Q} \bar{Z}_k^T \hat{E} \left[ \delta g \hat{Q} \bar{Z}_k H \right]$$

$$\hat{R}^t(H) = \hat{R} \circ \hat{R}^{t-1}(H) \quad \text{for } t \geq 2$$

where  $g(X) = E(H_1 | X) f(X)$  and  $\delta g = \frac{g(x) - \hat{g}(X)}{\hat{g}(X)} = \frac{Q^2 f - \hat{Q}^2 \hat{f}}{\hat{Q}^2 \hat{f}}$ .

Then

$$\begin{aligned}
& (-1)^{m-1} E B_m \\
&= \left\{ E \left[ Q^2 \left( \frac{B-\hat{B}}{\dot{B}} \right) \bar{Z}_k^T \right] \left\{ \hat{E} \left[ \delta g \hat{Q}^2 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{m-2} \times \right. \\
&\quad \left. \hat{E} \left[ \delta g \hat{Q} \bar{Z}_k \frac{\hat{Q}}{\hat{Q}} \Pi \left( Q \bar{Z}_k \left( \frac{P-\hat{P}}{\dot{P}} \right) \mid (Q \bar{Z}_k) \right) \right] \right\} \\
&= E \left[ Q \left( \frac{B-\hat{B}}{\dot{B}} \right) \frac{Q}{\hat{Q}} \hat{R}^{m-1} \left( \frac{\hat{Q}}{\hat{Q}} \Pi \left( Q \left( \frac{P-\hat{P}}{\dot{P}} \right) \mid (Q \bar{Z}_k) \right) \right) \right] \\
&= \hat{E} \left[ Q \left( \frac{B-\hat{B}}{\dot{B}} \right) \frac{Q f}{\hat{Q} \hat{f}} \hat{R}^{m-1} \left( \frac{\hat{Q}}{\hat{Q}} \Pi \left( Q \left( \frac{P-\hat{P}}{\dot{P}} \right) \mid (Q \bar{Z}_k) \right) \right) \right] \\
&\leq \left\{ \hat{E} \left[ Q \left( \frac{B-\hat{B}}{\dot{B}} \right) \frac{Q f}{\hat{Q} \hat{f}} \right]^2 \right\}^{\frac{1}{2}} \left\{ \hat{E} \left[ \hat{R}^{m-1} \left( \frac{\hat{Q}}{\hat{Q}} \Pi \left( Q \left( \frac{P-\hat{P}}{\dot{P}} \right) \mid (Q \bar{Z}_k) \right) \right) \right]^2 \right\}^{\frac{1}{2}}
\end{aligned}$$

by Cauchy Shwartz.

Now

$$\begin{aligned}
& \hat{E} \left[ \hat{R}^{m-1} \left( \frac{\hat{Q}}{\hat{Q}} \Pi \left( Q \left( \frac{P-\hat{P}}{\dot{P}} \right) \mid (Q \bar{Z}_k) \right) \right) \right]^2 \\
&\leq \hat{E} \left[ (\delta g)^2 \left[ \hat{R}^{m-2} \left( \frac{\hat{Q}}{\hat{Q}} \Pi \left( Q \left( \frac{P-\hat{P}}{\dot{P}} \right) \mid (Q \bar{Z}_k) \right) \right) \right]^2 \right]
\end{aligned}$$

by the projection operator having operator norm equal to 1

$$\leq \|\delta g\|_\infty^2 \hat{E} \left[ \hat{R}^{m-2} \left( \frac{\hat{Q}}{\hat{Q}} \Pi \left( Q \left( \frac{P-\hat{P}}{\dot{P}} \right) \mid (Q \bar{Z}_k) \right) \right) \right]^2$$

Iterating this calculation  $m - 1$  times we find

$$\begin{aligned}
& \widehat{E} \left[ \widehat{R}^{m-1} \left( \frac{\widehat{Q}}{\widehat{Q}} \Pi \left( Q \left( \frac{P - \widehat{P}}{\dot{P}} \right) | (Q \overline{Z}_k) \right) \right) \right]^2 \\
& \leq \|\delta g\|_\infty^{2(m-1)} \widehat{E} \left[ \frac{\widehat{Q}}{\widehat{Q}} \Pi \left( Q \left( \frac{P - \widehat{P}}{\dot{P}} \right) | (Q \overline{Z}_k) \right) \right]^2 \\
& \leq \|\delta g\|_\infty^{2(m-1)} \left\| \frac{\widehat{G}}{\widehat{G}} \right\|_\infty E \left( \Pi \left( Q \left( \frac{P - \widehat{P}}{\dot{P}} \right) | (Q \overline{Z}_k) \right) \right)^2 \\
& \leq \|\delta g\|_\infty^{2(m-1)} \left\| \frac{\widehat{G}}{\widehat{G}} \right\|_\infty \int Q^2 \left( \frac{P - \widehat{P}}{\dot{P}} \right)^2 f(X) dX \\
& \leq \|\delta g\|_\infty^{2(m-1)} \left\| \frac{\widehat{G}}{\widehat{G}} \right\|_\infty \left\| \frac{Q^2 f}{\dot{P}^2} \right\|_\infty \int (p(X) - \widehat{p}(X))^2 dX \\
& = \|\delta g\|_\infty^{2(m-1)} \left\| \frac{Q^2 f}{\dot{P}^2} \right\|_\infty \left[ \int (p(X) - \widehat{p}(X))^2 dX \right] (1 + o_p(1))
\end{aligned}$$

by  $\frac{\widehat{G}}{\widehat{G}} = (1 + o_p(1))$ . Next

$$\begin{aligned}
& \widehat{E} \left[ Q \left( \frac{B - \widehat{B}}{\dot{B}} \right) \frac{Qf}{\widehat{Q}\widehat{f}} \right]^2 \\
& = \int \frac{Q^2 f}{\dot{B}^2} \frac{G}{\widehat{G}} (b(X) - \widehat{b}(X))^2 dX \\
& \leq \left\| \frac{Q^2 f}{\dot{B}^2} \right\|_\infty \left\| \frac{G}{\widehat{G}} \right\|_\infty \int (b(X) - \widehat{b}(X))^2 dX \\
& = \left\| \frac{Q^2 f}{\dot{B}^2} \right\|_\infty \left[ \int (b(X) - \widehat{b}(X))^2 dX \right] (1 + o_p(1))
\end{aligned}$$

Then we know

$$|EB_m| \leq \left\{ \|\delta g\|_\infty^{m-1} \left\| \left( \frac{\dot{B}}{\dot{P}} G \right) \right\|_\infty^{\frac{1}{2}} \left\| \frac{\dot{P}}{\dot{B}} G \right\|_\infty^{\frac{1}{2}} (1 + o_p(1)) \times \left\{ \int (p(X) - \widehat{p}(X))^2 dX \right\}^{\frac{1}{2}} \left\{ \int (b(X) - \widehat{b}(X))^2 dX \right\}^{\frac{1}{2}} \right\}$$

■

To prove theorem 26, we first give the following univariate result. Suppose that we have a set

$$\left\{ \overline{\phi}_{k_j}^{k_{j+1}}(X), j = 0, 1, \dots, 2m \right\}$$

such that for each  $(k_j, k_{j+1})$  pair,  $\overline{\phi}_{k_j}^{k_{j+1}}(X)$  either spans  $\mathcal{V}_{\log_2(k^*)}$  or spans  $\bigoplus_{v=\log_2(k_j-1)}^{\log_2(k_{j+1})-1} \mathcal{W}_v$  where  $\log_2(k_{j+1})-1$  and  $\log_2(k_j-1)$  are both nonnegative integers and  $\log_2(k_j-1) \leq \log_2(k_{j+1})-1$ . Let  $K_{(k_j, k_{j+1})}(X, Y) \equiv \overline{\phi}_{k_j}^{k_{j+1}}(X)^T \overline{\phi}_{k_j}^{k_{j+1}}(Y)$ , we first introduce an important preliminary lemma.

**Lemma 52** *For  $m \geq 0$  and  $1 \leq j \leq m+1$ , if  $k^* \leq k_{2j-1} \asymp k^*$ , then  $k_{2j-2} = 1$ ; otherwise if  $k_{2j-1} \gg k^*$ , then  $\log_2(k_{2j-2}-1)$  and  $\log_2(k_{2j-1})-1$  are both nonnegative integers and  $k_{2j-2} = o(k_{2j-1})$ . Thus,*

$$\left( E \left[ \prod_{j=1}^{m+1} K_{(k_{2j-2}, k_{2j-1})}(X, X) \right] \right)^{-1} (\prod_{j=1}^{m+1} k_{2j-1}) = O(1)$$

**Proof.** (Lemma 52) case 1:  $m = 0$ ,

$$\begin{aligned} & E \left[ K_{(k_0, k_1)}(X, X) \right] \\ &= E \left[ \overline{\phi}_{k_0}^{k_1}(X)^T \overline{\phi}_{k_0}^{k_1}(X) \right] \\ &= k_1 - k_0 \end{aligned}$$



which follows from the orthonormality of wavelets.

case 2:  $m = 1$ . If  $k_1 \asymp k_3 \asymp k^*$ , then,

$$\begin{aligned} & E \left[ K_{(k_0, k_1)}(X, X) K_{(k_2, k_3)}(X, X) \right] \\ & \geq E \left[ \left( K_{(1, k^*)}(X, X) \right)^2 \right] \geq \left( E \left[ K_{(1, k^*)}(X, X) \right] \right)^2 = (k^*)^2 \asymp k_1 k_3. \end{aligned}$$

Similarly, if  $k_1 \asymp k_3 \gg k^*$ , and we further assume  $k_1 < k_3$  WLOG, then

$$\begin{aligned} & E \left[ K_{(k_0, k_1)}(X, X) K_{(k_2, k_3)}(X, X) \right] \\ & \geq E \left[ \left( K_{(\frac{k_1}{2}+1, k_1)}(X, X) \right)^2 \right] \geq \left( E \left[ K_{(\frac{k_1}{2}+1, k_1)}(X, X) \right] \right)^2 \\ & = \left( \frac{k_1}{2} \right)^2 \asymp k_1 k_3; \end{aligned}$$

otherwise, WLOG, we assume that  $k_1 = o(k_3)$ , then

$$\begin{aligned} & E \left[ K_{(k_0, k_1)}(X, X) K_{(k_2, k_3)}(X, X) \right] \\ & = E \left[ \sum_{r=k_0}^{k_1} \phi_r^2(X) \sum_{s=k_2}^{k_3} \phi_s^2(X) \right] \\ & \geq E \left[ \sum_{r=k_0}^{k_1} \phi_r^2(X) \sum_{s=\frac{k_3}{2}+1}^{k_3} \phi_s^2(X) \right] \end{aligned}$$

Here we further assume that  $k_1 = k^*$ . The proof for the case when  $k_1 > k^*$  follows similarly.

Let  $\phi^w(x)$  indicate the corresponding compactly supported father/mother wavelet on  $[L_w, U_w]$ , whose absolute value is bounded above by  $M_w$ . By the continuity of  $\phi^w(x)$ , there exists a set  $A$ , which is a union of finite number of disjoint open intervals,

such that  $|\phi^w(x)|$  is greater than  $\sqrt{\frac{1}{2(U_w - L_w)}}$  on  $A$ . moreover, the Lebesgue measure (i.e., the length) of  $A$  is greater than  $\frac{1}{2M_w^2}$  because  $\phi^w(x)$  is bounded and has unit length. Specifically,

$$\begin{aligned} 1 &= \int (\phi^w(x))^2 dx \\ &= \int_{|\phi^w(x)|^2 \leq \frac{1}{2(U_w - L_w)}} (\phi^w(x))^2 dx + \int_{|\phi^w(x)|^2 > \frac{1}{2(U_w - L_w)}} (\phi^w(x))^2 dx \\ &\leq \frac{1}{2(U_w - L_w)} (U_w - L_w) + M_w^2 \mu(A). \end{aligned}$$

By definition  $\{\phi_r(X) : 1 \leq r \leq k^*\}$  is a sequence of level  $\log_2 k^*$  scaled and translated father wavelets on the unit interval  $[0, 1]$ , therefore it is obvious that the lebesgue measure of the set

$$\begin{aligned} \tilde{A} &\equiv \left\{ x : \sum_{r=1}^{k^*} \phi_r^2(x) > \frac{k^*}{2(U_w - L_w)} \right\} \\ &\supset \bigcup_{r=1}^{k^*} \left\{ x : \frac{1}{k^*} \phi_r^2(x) > \frac{1}{2(U_w - L_w)} \right\} \end{aligned}$$

is greater than  $\frac{1}{2M_w^2(U_w - L_w)}$ . Furthermore, for  $1 \leq r \leq k^*$ , the set  $\left\{ x : \frac{1}{k^*} \phi_r^2(x) > \frac{1}{2(U_w - L_w)} \right\}$  consists of multiple disjoint open intervals whose lengths are all of order  $\frac{1}{k^*}$ . In contrast, the support for any level  $\log_2 k_3 - 1$  scaled and translated mother wavelet  $\phi_r$ ,  $\frac{k_3}{2} + 1 \leq r \leq k_3$ , is of order  $k_3^{-1} \ll (k^*)^{-1}$  as  $k^* = o(k_3)$ . Therefore, at least  $\frac{1}{2} \mu(\tilde{A})$  proportion of the level  $\log_2 k_3 - 1$  scaled and translated mother wavelets  $\left( \phi_{\frac{k_3}{2}+1}, \dots, \phi_{k_3} \right)$

have their support inside  $\tilde{A}$ . Hence,

$$\begin{aligned}
& E \left[ K_{(k_0, k_1)}(X, X) K_{(k_2, k_3)}(X, X) \right] \\
& > E \left[ 1_{\tilde{A}} \sum_{r=1}^{k^*} \phi_r^2(X) \sum_{s=\frac{k_3}{2}+1}^{k_3} \phi_s^2(X) \right] \\
& \geq \frac{k^*}{2(U_w - L_w)} E \left[ 1_{\tilde{A}} \sum_{s=\frac{k_3}{2}+1}^{k_3} \phi_s^2(X) \right] \\
& > \frac{k^*}{2(U_w - L_w)} \frac{1}{4M_w^2(U_w - L_w)} \frac{k_3}{2} \\
& \asymp k_1 k_3.
\end{aligned}$$

case 3:  $m > 1$ . WLOG, we assume that  $k_1 \leq k_3 \dots \leq k_{2m+1}$  and  $k_1 \asymp k_3 \dots \asymp k_{2l_1-1} \ll k_{2l_1+1} \asymp \dots \asymp k_{2l_2-1} \ll \dots \ll k_{2l_{t-1}+1} \asymp \dots \asymp k_{2l_t-1} = k_{2m+1}$  for  $1 \leq t \leq m+1$ ,  $0 = l_0 < l_1 \dots < l_t = m+1$ , then

$$\begin{aligned}
& E \left[ \prod_{j=1}^{m+1} K_{(k_{2j-2}, k_{2j-1})}(X, X) \right] \\
& \geq E \left[ \left( K_{(k_0, k_1)} \right)^{l_1} \prod_{r=2}^t \left( K_{\left( \frac{k_{2l_{r-1}+1}}{2} + 1, k_{2l_{r-1}+1} \right)} \right)^{l_r - l_{r-1}} \right] \\
& > E \left[ \left( \sum_{r=k_0}^{k_1} \phi_r^2(X) \right)^{l_1} \prod_{r=2}^t \left( \sum_{r=\frac{k_{2l_{r-1}+1}}{2}+1}^{k_{2l_{r-1}+1}} \phi_r^2(X) \right)^{l_r - l_{r-1}} \right]
\end{aligned}$$

By a similar argument as that in case 2, we can show that there exists a set  $\tilde{A}_1$ , which consists of multiple disjoint open intervals, such that  $\sum_{r=k_0}^{k_1} \phi_r^2(x) > c_1^\phi k_1$ ,  $\forall x \in \tilde{A}_1$  and  $\mu(\tilde{A}_1) > \delta_1^\phi$  for some positive constants  $c_1^\phi, \delta_1^\phi$ . moreover, by the multiresolution analysis (MRA) property of compactly supported wavelets and the fact that

$k_1 = o(k_{2l_1+1})$ , it is obvious that at least  $\frac{1}{2}\delta_1^\phi$  proportion of the level  $\log_2(k_{2l_1+1}) - 1$  scaled and translated mother wavelets  $\left\{\phi_{\frac{k_{2l_1+1}}{2}+1}, \dots, \phi_{k_{2l_1+1}}\right\}$  have support inside  $\tilde{A}_1$ . Hence we can find a set  $\tilde{A}_2 \subset \tilde{A}_1$ , which is also a union of disjoint open intervals, such that  $\sum_{r=\frac{k_{2l_1+1}}{2}+1}^{k_{2l_1+1}} \phi_r^2(x) > c_2^\phi k_{2l_1+1}$ ,  $\forall x \in \tilde{A}_2$  and  $\mu(\tilde{A}_2) > \delta_2^\phi \frac{\delta_1^\phi}{2}$  for some positive constants  $c_2^\phi, \delta_2^\phi$ . Furthermore, by applying this algorithm in a nested fashion  $t-1$  times, we can find a decreasing sequence of sets  $\tilde{A}_1 \supset \tilde{A}_2 \supset \dots \supset \tilde{A}_{t-2} \supset \tilde{A}_{t-1}$ , such that for any  $1 \leq s \leq t-2$ ,  $\sum_{r=\frac{k_{2l_{s+1}}}{2}+1}^{k_{2l_{s+1}}} \phi_r^2(x) > c_{s+1}^\phi k_{2l_{s+1}}$ ,  $\forall x \in \tilde{A}_{s+1}$ , and  $\mu(\tilde{A}_{s+1}) > \delta_{s+1}^\phi \prod_{r=1}^s \frac{\delta_r^\phi}{2}$  for some positive constants  $\{c_{s+1}^\phi, 1 \leq s \leq t-2\}$  and  $\{\delta_r^\phi, 1 \leq r \leq t-1\}$ . In addition,  $\prod_{r=1}^{t-1} \frac{\delta_r^\phi}{2}$  proportion of the level  $\log_2 k_{2l_{t-1}+1} - 1$  scaled and translated mother wavelets  $\left\{\phi_{\frac{k_{2l_{t-1}+1}}{2}+1}, \dots, \phi_{k_{2l_{t-1}+1}}\right\}$  have support inside  $\tilde{A}_{t-1}$ .

Hence,

$$\begin{aligned}
& E \left[ \prod_{j=1}^{m+1} K_{(k_{2j-2}, k_{2j-1})} (X, X) \right] \\
& > \left( \prod_{r=1}^{t-1} k_{2l_{r-1}+1}^{l_r - l_{r-1}} \right) E \left[ \left( 1_{A_{t-1}} \sum_{r=\frac{k_{2l_{t-1}+1}}{2}+1}^{k_{2l_{t-1}+1}} \phi_r^2(X) \right)^{l_t - l_{t-1}} \right] \\
& \geq \left( \prod_{r=1}^{t-1} k_{2l_{r-1}+1}^{l_r - l_{r-1}} \right) E \left[ \left( 1_{A_{t-1}} \sum_{r=\frac{k_{2l_{t-1}+1}}{2}+1}^{k_{2l_{t-1}+1}} \phi_r^2(X) \right)^{l_t - l_{t-1}} \right] \\
& > \left( \prod_{r=1}^{t-1} k_{2l_{r-1}+1}^{l_r - l_{r-1}} \right) \left( \frac{k_{2l_{t-1}+1}}{2} \prod_{r=1}^{t-1} \frac{\delta_r^\phi}{2} \right)^{l_t - l_{t-1}} \times \prod_{j=1}^{m+1} k_{2j-1}
\end{aligned}$$

■

Now, we are ready to prove the variance results for the univariate case.

**Lemma 53** *Suppose the assumptions in Lemma 52 hold, then*

$$\begin{aligned}
& \left\| \overline{\phi}_{k_0}^{k_1}(X_{i_1})^T \prod_{j=1}^m \left\{ \overline{\phi}_{k_{2j-2}}^{k_{2j-1}}(X_{i_{j+1}}) \overline{\phi}_{k_{2j}}^{k_{2j+1}}(X_{i_{j+1}})^T \right\} \overline{\phi}_{k_{2m}}^{k_{2m+1}}(X_{i_{m+2}}) \right\|^2 \\
&= \left\| \prod_{j=1}^{m+1} K_{(k_{2j-2}, k_{2j-1})}(X_{i_j}, X_{i_{j+1}}) \right\|^2 \\
&\asymp \prod_{j=1}^{m+1} k_{2j-1}
\end{aligned}$$

**Proof.** (Lemma: 53) case 1:  $m = 0$ .

$$\begin{aligned}
& E \left[ \overline{\phi}_{k_0}^{k_1}(X_{i_1})^T \overline{\phi}_{k_0}^{k_1}(X_{i_2}) \right]^2 \\
&= E \left[ \left( K_{(k_0, k_1)}(X_{i_1}, X_{i_2}) \right)^2 \right] \\
&= E \left[ K_{(k_0, k_1)}(X_{i_1}, X_{i_1}) \right] \\
&= k_1 - k_0
\end{aligned}$$

which follows from the orthonormality of wavelets.

Next, we prove the lemma for the case where  $m \geq 1$  in two steps.

a). We first prove that

$$\left\| \prod_{j=1}^{m+1} K_{(k_{2j-2}, k_{2j-1})}(X_{i_j}, X_{i_{j+1}}) \right\|^2 = O \left( \prod_{j=1}^{m+1} k_{2j-1} \right)$$

Consider the case  $m = 1$

$$\begin{aligned}
& E \left[ \left( K_{(k_0, k_1)} (X_{i_1}, X_{i_2}) K_{(k_2, k_3)} (X_{i_2}, X_{i_3}) \right)^2 \right] \\
&= E \left[ \left( K_{(k_0, k_1)} (X_{i_2}, X_{i_2}) K_{(k_2, k_3)} (X_{i_2}, X_{i_2}) \right) \right] \\
&\leq \|K_{(k_0, k_1)} (X, X)\|_\infty \|K_{(k_2, k_3)} (X, X)\|_\infty
\end{aligned}$$

if  $(k_0, k_1) = (1, k^*)$ , then  $\|K_{(k_0, k_1)} (X, X)\|_\infty = O(\|\phi_1^2(X)\|_\infty) = O(k_1 - k_0)$ . Similarly, if  $k_0 > k^*$  and  $\log_2(k_0 - 1) = \log_2(k_1) - 1$ , then  $\|K_{(k_0, k_1)} (X, X)\|_\infty = O(\|\phi_{k_0}^2(X)\|_\infty) = O(k_1 - k_0)$ ; otherwise, if  $k_0 > k^*$  and  $\log_2(k_0 - 1) > \log_2(k_1) - 1$ , then

$$\begin{aligned}
\|K_{(k_0, k_1)} (X, X)\|_\infty &= O \left( \sum_{t=0}^{\text{int}(\log_2(\frac{k_1}{k_0}))} k_0 2^t \right) \\
&= O(k_1)
\end{aligned}$$

Finally, for  $m > 1$

$$\begin{aligned}
& E \left[ \left( \prod_{j=1}^{m+1} K_{(k_{2j-2}, k_{2j-1})} (X_{i_j}, X_{i_{j+1}}) \right)^2 \right] \\
&= E \left[ K_{(k_0, k_1)} (X_{i_1}, X_{i_2})^2 \left( \prod_{j=2}^m K_{(k_{2j-2}, k_{2j-1})} (X_{i_j}, X_{i_{j+1}}) \right)^2 K_{(k_{2m}, k_{2m+1})} (X_{i_{m+1}}, X_{i_{m+2}})^2 \right] \\
&= E \left[ K_{(k_0, k_1)} (X_{i_2}, X_{i_2}) \left( \prod_{j=2}^m K_{(k_{2j-2}, k_{2j-1})} (X_{i_j}, X_{i_{j+1}}) \right)^2 K_{(J_{2m}, k_{2m+1})} (X_{i_{m+2}}, X_{i_{m+2}}) \right] \\
&\leq \|K_{(k_0, k_1)} (X, X)\|_\infty E \left[ \left( \prod_{j=2}^m K_{(k_{2j-2}, k_{2j-1})} (X_{i_j}, X_{i_{j+1}}) \right)^2 \right] \|K_{(k_{2m}, k_{2m+1})} (X, X)\|_\infty \\
&= \|K_{(k_0, k_1)} (X, X)\|_\infty \|K_{(k_{2m}, k_{2m+1})} (X, X)\|_\infty E \left[ \left( \prod_{j=2}^m K_{(k_{2j-2}, k_{2j-1})} (X_{i_j}, X_{i_{j+1}}) \right)^2 \right] \\
&\leq \prod_{j=1}^{m+1} \|K_{(k_{2j-2}, k_{2j-1})} (X, X)\|_\infty \\
&= O(\Pi_{j=1}^{m+1} k_{2j-1})
\end{aligned}$$

b). In the second step, we prove that

$$\frac{\Pi_{j=1}^{m+1} k_{2j-1}}{\left\| \prod_{j=1}^{m+1} K_{(k_{2j-2}, k_{2j-1})} (X_{i_j}, X_{i_{j+1}}) \right\|^2} = O(1) \quad (67)$$

We shall first show that  $\left\| \prod_{j=1}^{m+1} K_{(1, k_{2j-1})} (X_{i_j}, X_{i_{j+1}}) \right\|^2 \asymp \prod_{j=1}^{m+1} k_{2j-1}$ , and then complete the proof by showing that

$$\left\| \prod_{j=1}^{m+1} K_{(k_{2j-2}, k_{2j-1})} (X_{i_j}, X_{i_{j+1}}) \right\|^2 \asymp \left\| \prod_{j=1}^{m+1} K_{(1, k_{2j-1})} (X_{i_j}, X_{i_{j+1}}) \right\|^2 \quad (68)$$

Specifically,

$$\begin{aligned}
& \left\| \prod_{j=1}^{m+1} K_{(1,k_{2j-1})} (X_{i_j}, X_{i_{j+1}}) \right\|^2 \\
&= E \left[ K_{(1,k_1)} (X_{i_2}, X_{i_2}) \left( K_{(1,k_3)} (X_{i_2}, X_{i_3}) \right)^2 \prod_{j=3}^{m+1} \left( K_{(1,k_{2j-1})} (X_{i_j}, X_{i_{j+1}}) \right)^2 \right] \\
&= E \left[ K_{(1,k_1)} (X_{i_3}, X_{i_3}) K_{(1,k_3)} (X_{i_3}, X_{i_3}) \prod_{j=3}^{m+1} \left( K_{(1,k_{2j-1})} (X_{i_j}, X_{i_{j+1}}) \right)^2 \right] \\
&- E \left[ k_1 k_3 \{ h_{X_{i_3}} (X_{i_3}) - (\mathcal{K}_{(1,k_3)} \circ h_{X_{i_3}}) (X_{i_3}) \} \prod_{j=3}^{m+1} \left( K_{(1,k_{2j-1})} (X_{i_j}, X_{i_{j+1}}) \right)^2 \right]
\end{aligned}$$

where  $h_{X_{i_3}} (x) = \frac{1}{k_1 k_3} K_{(1,k_1)} (x, x) K_{(1,k_3)} (x, X_{i_3})$  and  $(\mathcal{K}_{(1,k_3)} \circ h) (\cdot) = E [h(X) K_{(1,k_3)} (X, \cdot)]$ .

As previously shown, there exists a positive constant  $C_\phi$  such that  $\sup_{x, X_{i_3}} |h_{X_{i_3}} (x)| \leq C_\phi$ . By the regularity property of compactly supported wavelets and the approximation property of the kernel  $K_{(1,k_3)} (\cdot, \cdot)$ , it is obvious that  $\|h_{X_{i_3}} (X_{i_2}) - (\mathcal{K}_{(1,k_3)} \circ h_{X_{i_3}}) (X_{i_3})\|_\infty = o(1)$ . Thus,

$$\begin{aligned}
& E \left[ k_1 k_3 \{ h_{X_{i_3}} (X_{i_2}) - (\mathcal{K}_{(1,k_3)} \circ h_{X_{i_3}}) (X_{i_3}) \} \right. \\
& \quad \left. \times \prod_{j=3}^{m+1} \left( K_{(1,k_{2j-1})} (X_{i_j}, X_{i_{j+1}}) \right)^2 \right] \\
&= o \left( \prod_{j=1}^{m+1} k_{2j-1} \right)
\end{aligned}$$



In fact, by arguing similarly as above,

$$\begin{aligned}
& \left\| \prod_{j=1}^{m+1} K_{(1,k_{2j-1})} (X_{i_j}, X_{i_{j+1}}) \right\|^2 \\
&= E \left[ \left\{ \begin{aligned} & K_{(1,k_1)} (X_{i_3}, X_{i_3}) K_{(1,k_3)} (X_{i_3}, X_{i_3}) \\ & \times \prod_{j=3}^{m+1} (K_{(1,k_{2j-1})} (X_{i_j}, X_{i_{j+1}}))^2 \end{aligned} \right\} \right] + o(\Pi_{j=1}^{m+1} k_{2j-1}) \\
&= E \left[ \left\{ \begin{aligned} & K_{(1,k_1)} (X_{i_4}, X_{i_4}) K_{(1,k_3)} (X_{i_4}, X_{i_4}) K_{(1,k_5)} (X_{i_4}, X_{i_4}) \\ & \times \prod_{j=4}^{m+1} (K_{(1,k_{2j-1})} (X_{i_j}, X_{i_{j+1}}))^2 \end{aligned} \right\} \right] \\
&+ o(\Pi_{j=1}^{m+1} k_{2j-1}) \\
&= E \left[ \prod_{j=1}^{m+1} K_{(1,k_{2j-1})} (X, X) \right] + o(\Pi_{j=1}^{m+1} k_{2j-1}) \\
&\asymp \prod_{j=1}^{m+1} k_{2j-1}
\end{aligned}$$

Now, we prove eq. (68).

$$\begin{aligned}
& \prod_{j=1}^{m+1} K_{(k_{2j-2}, k_{2j-1})} (X_{i_j}, X_{i_{j+1}}) \\
&= K_{(1,k_1)} (X_{i_1}, X_{i_2}) \prod_{j=2}^{m+1} K_{(k_{2j-2}, k_{2j-1})} (X_{i_j}, X_{i_{j+1}}) \\
&- K_{(1,k_0)} (X_{i_1}, X_{i_2}) \prod_{j=2}^{m+1} K_{(k_{2j-2}, k_{2j-1})} (X_{i_j}, X_{i_{j+1}}) \\
&=
\end{aligned}$$

$$\begin{aligned}
& K_{(1,k_1)}(X_{i_1}, X_{i_2}) \{K_{(1,k_3)}(X_{i_2}, X_{i_3}) - K_{(1,k_2)}(X_{i_2}, X_{i_3})\} \\
& \times \prod_{j=2}^{m+1} K_{(k_{2j-2}, k_{2j-1})}(X_{i_j}, X_{i_{j+1}}) \\
& - K_{(1,k_0)}(X_{i_1}, X_{i_2}) \prod_{j=2}^{m+1} K_{(k_{2j-2}, k_{2j-1})}(X_{i_j}, X_{i_{j+1}}) \\
& = \prod_{j=1}^{m+1} K_{(1,k_{2j-1})}(X_{i_j}, X_{i_{j+1}}) - \sum_{\{(k_{2j-2}^*, k_{2j-1}^*)\}} \prod_{j=1}^{m+1} K_{(k_{2j-2}^*, k_{2j-1}^*)}(X_{i_j}, X_{i_{j+1}})
\end{aligned}$$

where  $(k_{2j-2}^*, k_{2j-1}^*)$  may be  $(1, k_{2j-2})$  or  $(k_{2j-2}, k_{2j-1})$ , but  $\prod_{j=1}^{m+1} k_{2j-1}^* = o\left(\prod_{j=1}^{m+1} k_{2j-1}\right)$ .

Therefore  $\left\|\prod_{j=1}^{m+1} K_{(k_{2j-2}, k_{2j-1})}(X_{i_j}, X_{i_{j+1}})\right\|^2 \asymp \left\|\prod_{j=1}^{m+1} K_{(1, k_{2j-1})}(X_{i_j}, X_{i_{j+1}})\right\|^2$  as a direct consequence of the previously proved result that  $\left\|K_{(k_{2j-2}^*, k_{2j-1}^*)}(X_{i_j}, X_{i_{j+1}})\right\|^2 = O\left(\prod_{j=1}^{m+1} k_{2j-1}^*\right) = o\left(\prod_{j=1}^{m+1} k_{2j-1}\right)$ . The proof is complete. ■

**Proof.** (Theorem 26) We note that since the linear span of  $\overline{\varphi}_1^{k_j}(X)$  equals  $\bigotimes_{1 \leq r \leq l} \mathcal{V}_{r, \log_2(k_{j,r})}$ , we have

$$\begin{aligned}
& K_{(1,k_j)}(X_{i_j}, X_{i_{j+1}}) \\
& = \sum_{\left\{ \begin{array}{l} t_1, \dots, t_l : \\ 1 \leq t_u \leq k_{j,u} \\ u = 1, \dots, l \end{array} \right\}} \prod_{u=1}^l \overline{\varphi}_{t_u}(X_{i_j}^u) \prod_{u'=1}^l \overline{\varphi}_{t_{u'}}(X_{i_{j+1}}^{u'})
\end{aligned}$$

So that

$$\begin{aligned}
& K_{(1,k_j)}(X_{i_j}, X_{i_{j+1}}) \\
&= \prod_{u=1}^l \sum_{1 \leq t_u \leq k_{j,u}} \bar{\varphi}_{t_u}(X_{i_j}^u) \bar{\varphi}_{t_u}(X_{i_{j+1}}^u) \\
&= \prod_{u=1}^l K_{(1,k_{j,u})}(X_{i_j}^u, X_{i_{j+1}}^u)
\end{aligned}$$

The remainder of the proof then follows from Lemma 53 . ■

**Proof.** (Theorem 28) In the proof of Theorem 28, the following lemma plays a central role. Note that expectations and probabilities remain conditional on  $\hat{\theta}$  even when it is suppressed in the notation. ■

**Lemma 54** *Let  $\widehat{IF}_{m,m,\tilde{\psi}_k(\cdot)}^{(s)}$  be the symmetric kernel of  $\widehat{\mathbb{IF}}_{m,m,\tilde{\psi}_k(\cdot)}$ , then for any  $m \geq 2$  and  $1 \leq m_1 < m_2$ ,*

$$\begin{aligned}
& Var_{\theta} \left[ \left( \widehat{\mathbb{IF}}_{m,m,\tilde{\psi}_k(\cdot)} \right)^2 \right] - Var_{\hat{\theta}} \left[ \left( \widehat{\mathbb{IF}}_{m,m,\tilde{\psi}_k(\cdot)} \right)^2 \right] \\
&= O \left( n^{-m} \left\{ E_{\theta} \left[ \left( \widehat{IF}_{m,m,\tilde{i}_m}^{(s)} \right)^2 \right] - E_{\hat{\theta}} \left[ \left( \widehat{IF}_{m,m,\tilde{i}_m}^{(s)} \right)^2 \right] \right\} \right) \\
&+ O \left( \frac{1}{n} \left\{ E_{\theta} \left[ \widehat{IF}_{m,m,\tilde{i}_m}^{(s)} \right] \right\}^2 \right) \\
&+ o \left( \max \left( \frac{1}{n}, \frac{k^{m-2}}{n^{m-1}} \right) \right)
\end{aligned} \tag{69}$$

and

$$\begin{aligned}
& Cov_{\theta} \left( \widehat{\mathbb{IF}}_{m_1,m_1,\tilde{\psi}_k(\cdot)}, \widehat{\mathbb{IF}}_{m_2,m_2,\tilde{\psi}_k(\cdot)} \right) \\
&= O \left( \frac{1}{n} E_{\theta} \left[ \widehat{IF}_{m_1,m_1,\tilde{i}_m}^{(s)} \right] E_{\theta} \left[ \widehat{IF}_{m_2,m_2,\tilde{i}_m}^{(s)} \right] \right) + o \left( \max \left( \frac{1}{n}, \frac{k^{m_1-2}}{n^{m_1-1}} \right) \right)
\end{aligned} \tag{70}$$

The proof of this lemma is delayed as some technical details are involved. We first use this lemma to prove Theorem 28.

**Proof.** (Theorem 28). By the degeneracy of  $\widehat{\mathbb{IF}}_{t,t,\tilde{\psi}_k(\cdot)}$  for any  $t \leq m$  under  $F(\cdot; \widehat{\theta})$ ,

$$\begin{aligned} & \frac{Var_{\theta} \left[ \widehat{\mathbb{IF}}_{m,\tilde{\psi}_k(\cdot)} | \widehat{\theta} \right] - Var_{\widehat{\theta}} \left[ \widehat{\mathbb{IF}}_{m,\tilde{\psi}_k(\cdot)} | \widehat{\theta} \right]}{Var_{\widehat{\theta}} \left[ \widehat{\mathbb{IF}}_{m,\tilde{\psi}_k(\cdot)} | \widehat{\theta} \right]} \\ &= \frac{\left\{ \sum_{t=1}^m \left( Var_{\theta} \left[ \widehat{\mathbb{IF}}_{t,t,\tilde{\psi}_k(\cdot)} \right] - Var_{\widehat{\theta}} \left[ \widehat{\mathbb{IF}}_{t,t,\tilde{\psi}_k(\cdot)} \right] \right) + 2 \sum_{1 \leq t_1 < t_2 \leq m} Cov_{\theta} \left[ \widehat{\mathbb{IF}}_{t_1,t_1,\tilde{\psi}_k(\cdot)}, \widehat{\mathbb{IF}}_{t_2,t_2,\tilde{\psi}_k(\cdot)} \right] \right\}}{Var_{\widehat{\theta}} \left[ \widehat{\mathbb{IF}}_{m,\tilde{\psi}_k(\cdot)} | \widehat{\theta} \right]}, \end{aligned}$$

which equals

$$\begin{aligned} & \frac{\sum_{t=1}^m n^{-t} \left( E_{\theta} \left[ \left( \widehat{IF}_{t,t,\tilde{i}_m}^{(s)} \right)^2 \right] - E_{\widehat{\theta}} \left[ \left( \widehat{IF}_{t,t,\tilde{i}_m}^{(s)} \right)^2 \right] \right)}{\sum_{t=1}^m n^{-t} E_{\widehat{\theta}} \left[ \left( \widehat{IF}_{t,t,\tilde{i}_m}^{(s)} \right)^2 \right]} (1 + o(1)) \\ &+ \frac{o \left( \max \left( \frac{1}{n}, \frac{k^{m-2}}{n^{m-1}} \right) \right)}{Var_{\widehat{\theta}} \left[ \widehat{\mathbb{IF}}_{m,\tilde{\psi}_k(\cdot)} | \widehat{\theta} \right]} \end{aligned}$$

by Lemma 54.

By assumption,  $\sup_{o \in \mathcal{O}} |f(o; \hat{\theta}) - f(o; \theta)| \rightarrow 0$  as  $\|\hat{\theta} - \theta\| \rightarrow 0$ , hence

$$\begin{aligned}
& \frac{E_{\theta} \left[ \left( \widehat{IF}_{t,t,\bar{i}_m}^{(s)} \right)^2 \right] - E_{\hat{\theta}} \left[ \left( \widehat{IF}_{t,t,\bar{i}_m}^{(s)} \right)^2 \right]}{E_{\hat{\theta}} \left[ \left( \widehat{IF}_{t,t,\bar{i}_m}^{(s)} \right)^2 \right]} \\
&= \frac{E_{\hat{\theta}} \left[ \left( \widehat{IF}_{t,t,\bar{i}_m}^{(s)} \right)^2 \left( \prod_{s=1}^t \frac{f(O_{i_s}; \theta)}{f(O_{i_s}; \hat{\theta})} - 1 \right) \right]}{E_{\hat{\theta}} \left[ \left( \widehat{IF}_{t,t,\bar{i}_m}^{(s)} \right)^2 \right]} \\
&\leq \sup_{\mathbf{o}_{i_t}} \left| \prod_{s=1}^t \frac{f(O_{i_s}; \theta)}{f(O_{i_s}; \hat{\theta})} - 1 \right| \\
&= O \left( \sup_{o \in \mathcal{O}} |f(o; \hat{\theta}) - f(o; \theta)| \right) = o(1).
\end{aligned}$$

Furthermore,  $\frac{o\left(\max\left(\frac{1}{n}, \frac{k^{m-2}}{n^{m-1}}\right)\right)}{\text{Var}_{\hat{\theta}}[\mathbb{IF}_{m,\tilde{\psi}_k(\cdot)}|\hat{\theta}]} = o(1)$  as well since  $\text{Var}_{\hat{\theta}}[\mathbb{IF}_{m,\tilde{\psi}_k(\cdot)}|\hat{\theta}] \times \max\left(\frac{1}{n}, \frac{k^{m-1}}{n^m}\right)$

from Theorem 26. Thus the proof of Theorem 28 is complete. ■

Before giving out the proof of Lemma 54, we first introduce two useful propositions.

**Proposition 55** *For any  $m \geq 2$  and  $1 \leq t \leq m$ ,*

$$E_{\theta} \left[ \left\{ E_{\theta} \left( \widehat{IF}_{m,m,\bar{i}_m} | \mathbf{O}_{-i_t} \right) \right\}^2 \right] = o(k^{m-2}) \quad (71)$$

**Proof.** As proved in Theorem 31, for any  $m \geq 2$ ,

$$\begin{aligned}
& \widehat{IF}_{m,m,\bar{i}_m} \\
&= (-1)^{j-1} \left\{ \left[ \left( H_1 \hat{P} + H_2 \right) \dot{B} \bar{Z}_k^T \right]_{i_1} \left[ \prod_{s=3}^m \left\{ \left( \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right)_{i_s} - I_{k \times k} \right\} \right] \right. \\
& \quad \left. \times \left[ \bar{Z}_k \left( H_1 \hat{B} + H_3 \right) \dot{P} \right]_{i_2} \right\}
\end{aligned}$$

We first consider the case when  $m = 2$ .

$$\begin{aligned}
& - E_\theta \left[ \widehat{IF}_{2,2,i_1 i_2} | O_{i_2} \right] \\
& = E_\theta \left[ Q^2 \left( \frac{P - \widehat{P}}{\dot{P}} \right) \overline{Z}_k^T \right] \left[ \overline{Z}_k \left( H_1 \widehat{B} + H_3 \right) \dot{P} \right]_{i_2} \\
& = \widehat{E} \left[ \frac{f(X)}{\widehat{f}(X)} \frac{Q^2}{\widehat{Q}} \left( \frac{P - \widehat{P}}{\dot{P}} \right) \widehat{Q} \overline{Z}_k^T \right] \left[ \frac{\widehat{Q}}{\widehat{Q}} \overline{Z}_k \left( H_1 \widehat{B} + H_3 \right) \dot{P} \right]_{i_2} \\
& = \widehat{\Pi} \left[ \left( \frac{f(X)}{\widehat{f}(X)} \frac{Q^2}{\widehat{Q}} \left( \frac{P - \widehat{P}}{\dot{P}} \right) \right) | \left( \widehat{Q} \overline{Z}_k \right)_{i_2} \right] \left( \frac{\left( H_1 \widehat{B} + H_3 \right) \dot{P}}{\widehat{Q}} \right)_{i_2} \\
& = \left\| P - \widehat{P} \right\|_2 \frac{T_c(O_{i_2})}{\left\| P - \widehat{P} \right\|_2} \left( \frac{\left( H_1 \widehat{B} + H_3 \right) \dot{P}}{\widehat{Q}} \right)_{i_2}
\end{aligned}$$

where

$$T_c(O) \equiv \widehat{\Pi} \left[ \left( \frac{f(X)}{\widehat{f}(X)} \frac{Q^2}{\widehat{Q}} \frac{P - \widehat{P}}{\dot{P}} \right) | \left( \widehat{Q} \overline{Z}_k \right) \right]$$

Since assumptions (21) – (23) and  $Ai) - Aiv)$  are satisfied, it is easy to show that  $E \left[ \left( \frac{T_c(O_{i_2})}{\left\| P - \widehat{P} \right\|_2} \left( \frac{\left( H_1 \widehat{B} + H_3 \right) \dot{P}}{\widehat{Q}} \right)_{i_2} \right)^2 \right] = O(1)$ , and thus

$$E \left[ \left( E_\theta \left[ \widehat{IF}_{2,2,i_1 i_2} | O_{i_2} \right] \right)^2 \right] = o(1).$$

Similarly, we can prove that  $E \left[ \left( E_\theta \left[ \widehat{IF}_{2,2,i_1 i_2} | O_{i_1} \right] \right)^2 \right] = o(1)$ .

Next, we proceed by induction. We assume eq. (71) holds for  $m - 1$  and prove it is also true for  $m$  by considering different values of  $t$ .

i) If  $t = 1$ , then

$$\begin{aligned}
& (-1)^{m-1} E_\theta \left( \widehat{IF}_{m,m,\bar{i}_m} | \mathbf{O}_{-i_1} \right) \\
&= E_\theta \left[ Q^2 \left( \frac{P - \hat{P}}{\dot{P}} \right) \bar{Z}_k^T \right] \left[ \prod_{s=3}^m \left\{ \begin{array}{c} \left( \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right)_{i_s} \\ -I_{k \times k} \end{array} \right\} \right] \\
&\times \left[ \bar{Z}_k \left( H_1 \hat{B} + H_3 \right) \dot{P} \right]_{i_2} \\
&= \left\{ \begin{array}{c} \hat{E} \left[ \frac{f(X)}{\hat{f}(X)} Q^2 \left( \frac{P - \hat{P}}{\dot{P}} \right) \bar{Z}_k^T \right] \left( \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right)_{i_3} \\ -E_\theta \left[ Q^2 \left( \frac{P - \hat{P}}{\dot{P}} \right) \bar{Z}_k^T \right] \end{array} \right\} \times \\
&\prod_{s=4}^m \left\{ \begin{array}{c} \left( \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right)_{i_s} \\ -I_{k \times k} \end{array} \right\} \left[ \bar{Z}_k \left( H_1 \hat{B} + H_3 \right) \dot{P} \right]_{i_2} \\
&= \\
&\left\{ \frac{\dot{P} \dot{B} H_1}{\hat{Q}} \hat{\Pi} \left[ \left( \frac{f(X)}{\hat{f}(X)} \frac{Q^2}{\hat{Q}} \left( \frac{P - \hat{P}}{\dot{P}} \right) \right) | \hat{Q} \bar{Z}_k \right] \bar{Z}_k^T \right\}_{i_3} \times \\
&\prod_{s=4}^m \left\{ \begin{array}{c} \left( \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right)_{i_s} \\ -I_{k \times k} \end{array} \right\} \left[ \bar{Z}_k \left( H_1 \hat{B} + H_3 \right) \dot{P} \right]_{i_2} \\
&- E_\theta \left( \widehat{IF}_{m-1,m-1,i_1 i_2 i_4 \dots i_m} | \mathbf{O}_{-i_1} \right) \\
&= \left( \frac{\dot{P} \dot{B} H_1}{\hat{Q}} \right)_{i_3} T_c(O_{i_3}) \bar{Z}_{k,i_3}^T \prod_{s=4}^m \left\{ \begin{array}{c} \left( \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right)_{i_s} \\ -I_{k \times k} \end{array} \right\} \\
&\times \left[ \bar{Z}_k \left( H_1 \hat{B} + H_3 \right) \dot{P} \right]_{i_2} - E_\theta \left( \widehat{IF}_{m-1,m-1,i_1 i_2 i_4 \dots i_m} | \mathbf{O}_{-i_1} \right)
\end{aligned}$$

From the fact that  $(a - b)^2 \leq 2(a^2 + b^2)$ , we have

$$\begin{aligned}
& E_\theta \left[ \left( E_\theta \left[ \widehat{IF}_{m,m,\bar{i}_m} \mid \mathbf{O}_{-i_1} \right] \right)^2 \right] \\
& \leq 2E \left[ \left( \begin{pmatrix} \left( \frac{\dot{P}\dot{B}H_1}{\bar{Q}} \right)_{i_3} T_c(O_{i_3}) \bar{Z}_{k,i_3}^T \times \\ \prod_{s=4}^m \left\{ \begin{pmatrix} \dot{P}\dot{B}H_1 \bar{Z}_k \bar{Z}_k^T \end{pmatrix}_{i_s} \\ -I_{k \times k} \end{pmatrix} \right\} \left[ \bar{Z}_k (H_1 \hat{B} + H_3) \dot{P} \right]_{i_2} \end{pmatrix} \right)^2 \right] \\
& + 2E \left[ \left( E_\theta \left[ \widehat{IF}_{m-1,m-1,i_1 i_2 i_4 \dots i_m} \mid \mathbf{O}_{-i_1} \right] \right)^2 \right]
\end{aligned}$$

From Theorem 26, it can be shown that

$$\begin{aligned}
& E \left[ \left( \begin{pmatrix} \left( \frac{\dot{P}\dot{B}H_1}{\bar{Q}} \right)_{i_3} T_c(O_{i_3}) \bar{Z}_{k,i_3}^T \times \\ \prod_{s=4}^m \left\{ \begin{pmatrix} \dot{P}\dot{B}H_1 \bar{Z}_k \bar{Z}_k^T \end{pmatrix}_{i_s} \\ -I_{k \times k} \end{pmatrix} \right\} \left[ \bar{Z}_k (H_1 \hat{B} + H_3) \dot{P} \right]_{i_2} \end{pmatrix} \right)^2 \right] \\
& = \left\| P - \hat{P} \right\|_\infty O(k^{m-2}) \\
& = o(k^{m-2})
\end{aligned}$$

By the induction assumption,  $E_\theta \left[ \left\{ E_\theta \left( \widehat{IF}_{m-1,m-1,i_1 i_2 i_4 \dots i_m} \mid O_{i_1} \right) \right\}^2 \right] = o(k^{m-3})$ .

Therefore eq. (71) holds when  $t = 1$ .

ii) Following the same argument as above, we can prove that eq. (71) also holds for  $t = 2$ .



iii) If  $3 \leq t \leq m$ , WLOG, assume  $t = 3$ , then

$$\begin{aligned}
& (-1)^{m-1} E \left[ \widehat{IF}_{m,m,\bar{i}_m} | \mathbf{O}_{-i_3} \right] \\
&= \left[ \left( H_1 \widehat{P} + H_2 \right) \dot{B} \bar{Z}_k^T \right]_{i_1} \left( E_\theta \left[ Q^2 \bar{Z}_k \bar{Z}_k^T \right] - I \right) \times \\
& \prod_{s=4}^m \left\{ \begin{array}{c} \left( \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right)_{i_s} \\ -I_{k \times k} \end{array} \right\} \left[ \bar{Z}_k \left( H_1 \widehat{B} + H_3 \right) \dot{P} \right]_{i_2} \\
&= \left[ \left( H_1 \widehat{P} + H_2 \right) \dot{B} \bar{Z}_k^T \right]_{i_1} \widehat{E} \left[ \delta g \widehat{Q}^2 \bar{Z}_k \bar{Z}_k^T \right] \left( \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right)_{i_4} \\
& \times \prod_{s=5}^m \left\{ \left( \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right)_{i_s} - I_{k \times k} \right\} \left[ \bar{Z}_k \left( H_1 \widehat{B} + H_3 \right) \dot{P} \right]_{i_2} \left( \equiv \widehat{T} \right) \\
& - E_\theta \left[ \widehat{IF}_{m-1,m-1,i_1 i_2 i_3 i_5 \dots i_m} | \mathbf{O}_{-i_3} \right]
\end{aligned}$$

moreover, it can be shown that  $E_\theta \left[ \widehat{T}^2 \right] = O \left( \|\delta g\|_\infty^2 k^{m-2} \right) = o(k^{m-2})$  following the proof of Theorem 26 but replacing  $K_k(X_{i_1}, X_{i_2})$  with  $K_k^\dagger(X_{i_1}, X_{i_2}) \equiv \bar{\phi}_0^k(X_{i_1})^T \widehat{E} \left[ \delta g \widehat{Q}^2 \bar{Z}_k \bar{Z}_k^T \right] \bar{\phi}_0^k(X_{i_2})$ . Specifically,

$$\begin{aligned}
& \int \left( K_k^\dagger(X_{i_1}, X_{i_2}) \right)^2 d\mu(O_{i_1}; \theta) \\
&= \text{tr} \left( \widehat{E} \left[ \delta g \widehat{Q}^2 \bar{Z}_k \bar{Z}_k^T \right] \bar{\phi}_0^k(X_{i_2}) \bar{\phi}_0^{k,T}(X_{i_2}) \widehat{E} \left[ \delta g \widehat{Q}^2 \bar{Z}_k \bar{Z}_k^T \right] \right) \\
&= \bar{\phi}_0^{k,T}(X_{i_2}) \left( \widehat{E} \left[ \delta g \widehat{Q}^2 \bar{Z}_k \bar{Z}_k^T \right] \right)^2 \bar{\phi}_0^k(X_{i_2}) \\
&\leq \|\delta g\|_\infty^2 \bar{\phi}_0^{k,T}(X_{i_2}) \bar{\phi}_0^k(X_{i_2})
\end{aligned}$$

The last inequality holds because  $\|\delta g\|_\infty I_{k \times k} - \widehat{E} \left[ \delta g \widehat{Q}^2 \bar{Z}_k \bar{Z}_k^T \right]$  is a semi-positive definite symmetric matrix.  $E_\theta \left[ \left( E_\theta \left[ \widehat{IF}_{m-1,m-1,i_1 i_2 i_3 i_5 \dots i_m} | \mathbf{O}_{-i_3} \right] \right)^2 \right]$  is of order  $o(k^{m-3})$  by induction assumption. Now, the proof of this proposition is complete. moreover,

by arguing similarly, the result above can be generalized in a straightforward manner to the following proposition. ■

**Proposition 56** *For any  $m \geq 2$  and  $1 \leq t < m$ ,*

$$E_{\theta} \left[ \left\{ E_{\theta} \left( \widehat{IF}_{m,m,\bar{i}_m} | O_{i_{s_1}}, \dots, O_{i_{s_t}} \right) \right\}^2 \right] = o(k^{t-1})$$

Finally, we are now ready to prove Lemma 54.

**Proof.** (Lemma 54) Throughout the proof, we repeatedly use the result that  $\widehat{\mathbb{IF}}_{m,m,\tilde{\psi}_k(\cdot)}$  for any  $m \geq 2$  is degenerate under  $F(\cdot; \hat{\theta})$ . We first prove eq. (69).

$$\begin{aligned} & Var_{\theta} \left[ \left( \widehat{\mathbb{IF}}_{m,m,\tilde{\psi}_k(\cdot)} \right)^2 \right] - Var_{\hat{\theta}} \left[ \left( \widehat{\mathbb{IF}}_{m,m,\tilde{\psi}_k(\cdot)} \right)^2 \right] \\ &= E_{\theta} \left[ \left( \widehat{\mathbb{IF}}_{m,m,\tilde{\psi}_k(\cdot)} \right)^2 \right] - E_{\hat{\theta}} \left[ \left( \widehat{\mathbb{IF}}_{m,m,\tilde{\psi}_k(\cdot)} \right)^2 \right] - \left( E_{\theta} \left[ \widehat{\mathbb{IF}}_{m,m,\tilde{\psi}_k(\cdot)} \right] \right)^2 \\ &= E_{\hat{\theta}} \left[ \left( \widehat{\mathbb{IF}}_{m,m,\tilde{\psi}_k(\cdot)} \right)^2 \left( \prod_{i=1}^n \frac{f(O_i)}{\hat{f}(O_i)} - 1 \right) \right] - \left( E_{\theta} \left[ \widehat{\mathbb{IF}}_{m,m,\tilde{\psi}_k(\cdot)} \right] \right)^2 \end{aligned}$$

can be written as a sum of four terms as below:

$$\begin{aligned}
& E_{\hat{\theta}} \left[ \left( \widehat{\mathbb{IF}}_{m,m,\tilde{\psi}_k(\cdot)} \right)^2 \left( \prod_{i=1}^n \frac{f(O_i)}{\widehat{f}(O_i)} - 1 \right) \right] - \left( E_{\theta} \left[ \widehat{\mathbb{IF}}_{m,m,\tilde{\psi}_k(\cdot)} \right] \right)^2 \\
&= E_{\hat{\theta}} \left\{ \left[ \binom{n}{m} \right]^{-2} \left( \sum_{i_1 < i_2 \dots < i_m} \widehat{IF}_{m,m,\bar{i}_m}^{(s)} \right) \right. \\
&\quad \times \left. \left( \sum_{r_1 < r_2 \dots < r_m} \widehat{IF}_{m,m,\bar{r}_m}^{(s)} \right) \left( \prod_{i=1}^n \frac{f(O_i)}{\widehat{f}(O_i)} - 1 \right) \right\} \\
&\quad - \left( E_{\theta} \left[ \widehat{\mathbb{IF}}_{m,m,\tilde{\psi}_k(\cdot)} \right] \right)^2 \\
&= E_{\hat{\theta}} \left\{ \left[ \binom{n}{m} \right]^{-2} \left[ \begin{aligned} & \sum_{i_1 < i_2 \dots < i_m} \left( \widehat{IF}_{m,m,\bar{i}_m}^{(s)} \right)^2 \\ & + \sum_{\bar{i}_m \cap \bar{r}_m = \emptyset} \widehat{IF}_{m,m,\bar{i}_m}^{(s)} \widehat{IF}_{m,m,\bar{r}_m}^{(s)} \\ & + \sum_{1 \leq \#(\bar{i}_m \cap \bar{r}_m) < m} \widehat{IF}_{m,m,\bar{i}_m}^{(s)} \widehat{IF}_{m,m,\bar{r}_m}^{(s)} \end{aligned} \right] \right. \\
&\quad \times \left. \left( \prod_{i=1}^n \frac{f(O_i)}{\widehat{f}(O_i)} - 1 \right) \right\} \\
&\quad - \left( E_{\theta} \left[ \widehat{\mathbb{IF}}_{m,m,\tilde{\psi}_k(\cdot)} \right] \right)^2
\end{aligned}$$

where  $\#(\bar{i}_m \cap \bar{r}_m)$  is the number of elements in the intersection set  $\{i_1, \dots, i_m\} \cap \{r_1, \dots, r_m\}$ .

The first term:

$$\begin{aligned}
& E_{\hat{\theta}} \left[ \left[ \binom{n}{m} \right]^{-2} \sum_{i_1 < i_2 \dots < i_m} \widehat{IF}_{m,m,\bar{i}_m}^{(s)2} \left( \prod_{i=1}^n \frac{f(O_i)}{\widehat{f}(O_i)} - 1 \right) \right] \\
&= \left[ \binom{n}{m} \right]^{-1} E_{\hat{\theta}} \left[ \widehat{IF}_{m,m,\bar{i}_m}^{(s)2} \left( \prod_{i=1}^n \frac{f(O_i)}{\widehat{f}(O_i)} - 1 \right) \right] \\
&= \left[ \binom{n}{m} \right]^{-1} \left( E_{\theta} \left[ \widehat{IF}_{m,m,\bar{i}_m}^{(s)2} \right] - E_{\hat{\theta}} \left[ \widehat{IF}_{m,m,\bar{i}_m}^{(s)2} \right] \right)
\end{aligned}$$

The second term:

$$\begin{aligned}
& E_{\hat{\theta}} \left[ \left[ \binom{n}{m} \right]^{-2} \left( \sum_{\bar{i}_m \cap \bar{r}_m = \emptyset} \widehat{IF}_{m,m,\bar{i}_m}^{(s)} \widehat{IF}_{m,m,\bar{r}_m}^{(s)} \right) \left( \prod_{i=1}^n \frac{f(O_i)}{\widehat{f}(O_i)} - 1 \right) \right] \\
&= \left[ \binom{n}{m} \right]^{-2} E_{\theta} \left[ \sum_{\bar{i}_m \cap \bar{r}_m = \emptyset} \widehat{IF}_{m,m,\bar{i}_m}^{(s)} \widehat{IF}_{m,m,\bar{r}_m}^{(s)} \right] \\
&= \left[ \binom{n}{m} \right]^{-2} \binom{n}{m} \binom{n-m}{m} E_{\theta} \left[ \widehat{IF}_{m,m,\bar{i}_m}^{(s)} \widehat{IF}_{m,m,\bar{r}_m}^{(s)} \right] \\
&= \frac{(n-m) \dots (n-2m+1)}{n(n-1) \dots (n-m+1)} \left( E_{\theta} \left[ \widehat{IF}_{m,m,\bar{i}_m}^{(s)} \right] \right)^2
\end{aligned}$$

Subtracting the fourth term from the second term, we have

$$\begin{aligned}
& \left[ \frac{(n-m) \dots (n-2m+1)}{n(n-1) \dots (n-m+1)} - 1 \right] \left( E_{\theta} \left[ \widehat{IF}_{m,m,\bar{i}_m}^{(s)} \right] \right)^2 \\
&= O \left( \frac{1}{n} \left( E_{\theta} \left[ \widehat{IF}_{m,m,\bar{i}_m}^{(s)} \right] \right)^2 \right)
\end{aligned}$$

The third term:

$$\begin{aligned}
& E_{\hat{\theta}} \left[ \left[ \binom{n}{m} \right]^{-2} \left( \sum_{1 \leq \#(\bar{i}_m \cap \bar{r}_m) < m} \widehat{IF}_{m,m,\bar{i}_m}^{(s)} \widehat{IF}_{m,m,\bar{r}_m}^{(s)} \right) \left( \prod_{i=1}^n \frac{f(O_i)}{\widehat{f}(O_i)} - 1 \right) \right] \\
&= \sum_{t=1}^{m-1} E_{\theta} \left[ \left[ \binom{n}{m} \right]^{-2} \left( \sum_{\#(\bar{i}_m \cap \bar{r}_m) = t} \widehat{IF}_{m,m,\bar{i}_m}^{(s)} \widehat{IF}_{m,m,\bar{r}_m}^{(s)} \right) \right] \\
&= \sum_{t=1}^{m-1} \left[ \binom{n}{m} \right]^{-2} \binom{n}{2m-t} \left[ \binom{m}{t} \right]^2 E_{\theta} \left[ \widehat{IF}_{m,m,i_1 i_2 \dots i_t i_{t+1} \dots i_m}^{(s)} \widehat{IF}_{m,m,i_1 i_2 \dots i_t i_{m+1} \dots i_{2m-t}}^{(s)} \right] \\
&= O \left( \sum_{t=1}^{m-1} n^{-t} E_{\theta} \left[ \left( E_{\theta} \left[ \widehat{IF}_{m,m,\bar{i}_m}^{(s)} | \mathbf{O}_{\bar{i}_t} \right] \right)^2 \right] \right) \\
&= O \left( \sum_{t=1}^{m-1} n^{-t} \left\{ E_{\theta} \left[ \left( \widehat{IF}_{m,m,\bar{i}_m} | O_{s_1}, \dots, O_{s_t} \right)^2 \right] E_{\theta} \left[ \left( \widehat{IF}_{m,m,\bar{i}_m} | O_{v_1}, \dots, O_{v_t} \right)^2 \right] \right\}^{1/2} \right) \\
&= o \left( \max \left( \frac{1}{n}, \frac{k^{m-2}}{n^{m-1}} \right) \right)
\end{aligned}$$

The last two equalities follow from Cauchy-Shwartz inequality and Lemma 54. Specifically,

$$\begin{aligned}
& E_\theta \left( E_\theta \left[ \widehat{IF}_{m,m,\bar{i}_m}^{(s)} | \mathbf{O}_{\bar{i}_t} \right] \right)^2 \\
&= E_\theta \left[ E_\theta \left( \widehat{IF}_{m,m,\bar{i}_m^*} | O_{i_1}, \dots, O_{i_t} \right) E_\theta \left( \widehat{IF}_{m,m,\bar{r}_m^*} | O_{i_1}, \dots, O_{i_t} \right) \right] \\
&\leq \left\{ E_\theta \left[ \left( E_\theta \left( \widehat{IF}_{m,m,\bar{i}_m^*} | O_{i_1}, \dots, O_{i_t} \right) \right)^2 \right] \right\}^{1/2} \\
&\times \left\{ E_\theta \left[ \left( E_\theta \left( \widehat{IF}_{m,m,\bar{r}_m^*} | O_{i_1}, \dots, O_{i_t} \right) \right)^2 \right] \right\}^{1/2}.
\end{aligned}$$

where  $\bar{i}_m^*$  and  $\bar{r}_m^*$  are two permutations of  $(i_1, i_2, \dots, i_m)$ .

Next, we prove eq. 70 for any  $1 \leq m_1 < m_2$ . Here we also rewrite  $Cov_\theta \left( \widehat{\mathbb{IF}}_{m_1, m_1, \tilde{\psi}_k(\cdot)}, \widehat{\mathbb{IF}}_{m_2, m_2, \tilde{\psi}_k(\cdot)} \right)$

as a sum of four terms.

$$\begin{aligned}
& Cov_\theta \left( \widehat{\mathbb{IF}}_{m_1, m_1, \tilde{\psi}_k(\cdot)}, \widehat{\mathbb{IF}}_{m_2, m_2, \tilde{\psi}_k(\cdot)} \right) \\
&= E_\theta \left[ \widehat{\mathbb{IF}}_{m_1, m_1, \tilde{\psi}_k(\cdot)} \widehat{\mathbb{IF}}_{m_2, m_2, \tilde{\psi}_k(\cdot)} \right] - E_\theta \left[ \widehat{IF}_{m_1, m_1, \bar{i}_m}^{(s)} \right] E_\theta \left[ \widehat{IF}_{m_2, m_2, \bar{i}_m}^{(s)} \right] \\
&= E_\theta \left[ \frac{1}{\binom{n}{m_1} \binom{n}{m_2}} \sum_{i_1 < i_2 \dots < i_{m_1}} \widehat{IF}_{m_1, m_1, \bar{i}_{m_1}}^{(s)} \sum_{r_1 < r_2 \dots < r_{m_2}} \widehat{IF}_{m_2, m_2, \bar{r}_{m_2}}^{(s)} \right] \\
&\quad - E_\theta \left[ \widehat{IF}_{m_1, m_1, \bar{i}_m}^{(s)} \right] E_\theta \left[ \widehat{IF}_{m_2, m_2, \bar{i}_m}^{(s)} \right] \\
&= E_\theta \left[ \frac{1}{\binom{n}{m_1} \binom{n}{m_2}} \left\{ \left( \sum_{\bar{i}_{m_1} \subset \bar{r}_{m_2}} + \sum_{\bar{i}_{m_1} \cap \bar{r}_{m_2} = \emptyset} + \sum_{1 \leq \#(\bar{i}_{m_1} \cap \bar{r}_{m_2}) < m_1} \right) \right. \right. \\
&\quad \left. \left. \times \widehat{IF}_{m_1, m_1, \bar{i}_{m_1}}^{(s)} \widehat{IF}_{m_2, m_2, \bar{r}_{m_2}}^{(s)} \right\} \right] \\
&\quad - E_\theta \left[ \widehat{IF}_{m_1, m_1, \bar{i}_m}^{(s)} \right] E_\theta \left[ \widehat{IF}_{m_2, m_2, \bar{i}_m}^{(s)} \right]
\end{aligned}$$

The first term:

$$\begin{aligned}
& E_\theta \left[ \frac{1}{\binom{n}{m_1} \binom{n}{m_2}} \sum_{\bar{i}_{m_1} \subset \bar{r}_{m_2}} \widehat{IF}_{m_1, m_1, \bar{i}_{m_1}}^{(s)} \widehat{IF}_{m_2, m_2, \bar{r}_{m_2}}^{(s)} \right] \\
&= \frac{1}{\binom{n}{m_1} \binom{n}{m_2}} \binom{n}{m_2} \binom{m_2}{m_1} E_\theta \left[ \widehat{IF}_{m_1, m_1, \bar{i}_{m_1}}^{(s)} \widehat{IF}_{m_2, m_2, \bar{r}_{m_2}}^{(s)} \right] \\
&= \frac{\binom{m_2}{m_1}}{\binom{n}{m_1}} E_\theta \left[ \widehat{IF}_{m_1, m_1, \bar{i}_{m_1}}^{(s)} E_\theta \left[ \widehat{IF}_{m_2, m_2, \bar{r}_{m_2}}^{(s)} | \mathbf{O}_{\bar{i}_{m_1}} \right] \right] \\
&\leq \frac{\binom{m_2}{m_1}}{\binom{n}{m_1}} \left\{ E_\theta \left[ \left( \widehat{IF}_{m_1, m_1, \bar{i}_{m_1}}^{(s)} \right)^2 \right] E_\theta \left[ \left( E_\theta \left[ \widehat{IF}_{m_2, m_2, \bar{r}_{m_2}}^{(s)} | \mathbf{O}_{\bar{i}_{m_1}} \right] \right)^2 \right] \right\}^{1/2} \\
&= o \left( \frac{k^{m_1-1}}{n^{m_1}} \right),
\end{aligned}$$

which follows from Cauchy-Schwartz inequality, Theorem (26), and Lemma 54.

The difference between the second and the fourth terms equals

$$\begin{aligned}
& E_\theta \left[ \frac{1}{\binom{n}{m_1} \binom{n}{m_2}} \sum_{\bar{i}_{m_1} \cap \bar{r}_{m_2} = \emptyset} \widehat{IF}_{m_1, m_1, \bar{i}_{m_1}}^{(s)} \widehat{IF}_{m_2, m_2, \bar{r}_{m_2}}^{(s)} \right] \\
&- E_\theta \left[ \widehat{IF}_{m_1, m_1, \bar{i}_m}^{(s)} \right] E_\theta \left[ \widehat{IF}_{m_2, m_2, \bar{i}_m}^{(s)} \right] \\
&= \left( \frac{1}{\binom{n}{m_1} \binom{n}{m_2}} \binom{n}{m_1} \binom{n-m_1}{m_2} - 1 \right) E_\theta \left[ \widehat{IF}_{m_1, m_1, \bar{i}_m}^{(s)} \right] E_\theta \left[ \widehat{IF}_{m_2, m_2, \bar{i}_m}^{(s)} \right] \\
&= O \left( \frac{1}{n} E_\theta \left[ \widehat{IF}_{m_1, m_1, \bar{i}_m}^{(s)} \right] E_\theta \left[ \widehat{IF}_{m_2, m_2, \bar{i}_m}^{(s)} \right] \right)
\end{aligned}$$

The third term:

$$\begin{aligned}
& E_\theta \left[ \frac{1}{\binom{n}{m_1} \binom{n}{m_2}} \sum_{1 \leq \#(\bar{i}_{m_1} \cap \bar{r}_{m_2}) < m_1} \widehat{IF}_{m_1, m_1, \bar{i}_{m_1}}^{(s)} \widehat{IF}_{m_2, m_2, \bar{r}_{m_2}}^{(s)} \right] \\
&= \sum_{t=1}^{m_1-1} \frac{1}{\binom{n}{m_1} \binom{n}{m_2}} E_\theta \left[ \sum_{\#(\bar{i}_{m_1} \cap \bar{r}_{m_2})=t} \widehat{IF}_{m_1, m_1, \bar{i}_{m_1}}^{(s)} \widehat{IF}_{m_2, m_2, \bar{r}_{m_2}}^{(s)} \right] \\
&= \sum_{t=1}^{m_1-1} \left\{ \frac{\binom{n}{m_2+m_1-t} \binom{m_2+m_1-t}{m_2} \binom{m_2}{t}}{\binom{n}{m_1} \binom{n}{m_2}} \times \right. \\
&\quad \left. E_\theta \left[ \widehat{IF}_{m_1, m_1, i_1 i_2 \dots i_t i_{t+1} \dots i_{m_1}}^{(s)} \widehat{IF}_{m_2, m_2, i_1 \dots i_t i_{m_1+1} \dots i_{m_1+m_2-t}}^{(s)} \right] \right\} \\
&= o \left( \max \left( \frac{1}{n}, \frac{k^{m_1-2}}{n^{m_1-1}} \right) \right),
\end{aligned}$$

which also follows from Cauchy-Schwartz inequality and Lemma 54. ■

**Proof.** (Eq. (36)) We prove the bias property of  $\widehat{\psi}_{m,k}^{\text{mod}}$  by induction.

For  $m = 2$ , The estimation bias is given by

$$\begin{aligned}
& - \left\{ E \left[ Q^2 \left( \frac{P - \widehat{P}}{\dot{P}} \right) \bar{Z}_k^T \right] E \left[ Q^2 \left( \frac{B - \widehat{B}}{\dot{B}} \right) \bar{Z}_k \right] \right\} \\
& + \left\{ E \left[ Q^2 \left( \frac{P - \widehat{P}}{\dot{P}} \right) \bar{Z}_k^T \right] \left\{ E \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} \right. \\
& \quad \left. \times E \left[ \bar{Z}_k Q^2 \left( \frac{B - \widehat{B}}{\dot{B}} \right) \right] \right\} \\
& =
\end{aligned}$$

$$\begin{aligned}
& \left\{ \begin{aligned} & E \left[ \left( H_1 \hat{P} + H_2 \right) \dot{B} \bar{Z}_k^T \right] \left[ \left\{ E \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} - I \right] \\ & \times E \left[ \bar{Z}_k \left( H_1 \hat{B} + H_3 \right) \dot{P} \right] \end{aligned} \right\} \\
& = - \left\{ \begin{aligned} & E \left[ \left( H_1 \hat{P} + H_2 \right) \dot{B} \bar{Z}_k^T \right] \left\{ E \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] - I \right\} \\ & \times E \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right]^{-1} E \left[ \bar{Z}_k \left( H_1 \hat{B} + H_3 \right) \dot{P} \right] \end{aligned} \right\}
\end{aligned}$$

Suppose the bias formula holds for  $m$ , then the bias at  $m + 1$  is

$$\begin{aligned}
& (-1)^{m-1} \left\{ \begin{aligned} & E \left[ Q^2 \left( \frac{P - \hat{P}}{\dot{P}} \right) \bar{Z}_k^T \right] \left\{ E \left[ Q^2 \bar{Z}_k \bar{Z}_k^T \right] - I \right\} \\ & \times \prod_{s=3}^m \left\{ \hat{E}_s \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} \left\{ E \left[ Q^2 \bar{Z}_k \bar{Z}_k^T \right] - \hat{E}_s \left[ Q^2 \bar{Z}_k \bar{Z}_k^T \right] \right\} \\ & \times \left\{ E \left[ Q^2 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} E \left[ Q^2 \bar{Z}_k \left( \frac{B - \hat{B}}{\dot{B}} \right) \right] \end{aligned} \right\} \\
& + (-1)^m E \left[ Q^2 \left( \frac{P - \hat{P}}{\dot{P}} \right) \bar{Z}_k^T \right] \left( E \left[ \left( \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right)_{i_2} \right] - I \right) \times \\
& \left[ \prod_{s=3}^m \left\{ \hat{E}_s \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} \left\{ E \left[ \left( \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right)_{i_s} \right] - \hat{E}_s \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\} \right] \\
& \times \left\{ \hat{E}_{m+1} \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} E \left[ \bar{Z}_k Q^2 \left( \frac{B - \hat{B}}{\dot{B}} \right) \right] \\
& =
\end{aligned}$$



$$\begin{aligned}
& (-1)^m \left\{ \begin{aligned} & E \left[ Q^2 \left( \frac{P-\hat{P}}{\hat{P}} \right) \bar{Z}_k^T \right] \left\{ E \left[ Q^2 \bar{Z}_k \bar{Z}_k^T \right] - I \right\} \\ & \times \prod_{s=3}^m \left\{ \hat{E}_s \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} \left\{ \begin{aligned} & E \left[ Q^2 \bar{Z}_k \bar{Z}_k^T \right] \\ & - \hat{E}_s \left[ Q^2 \bar{Z}_k \bar{Z}_k^T \right] \end{aligned} \right\} \\ & \times \left[ \left\{ \hat{E}_{m+1} \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} - \left\{ E \left[ Q^2 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} \right] \\ & \times E \left[ \bar{Z}_k Q^2 \left( \frac{B-\hat{B}}{\hat{B}} \right) \right] \end{aligned} \right\} \\
& = (-1)^m \left\{ \begin{aligned} & E \left[ Q^2 \left( \frac{P-\hat{P}}{\hat{P}} \right) \bar{Z}_k^T \right] \left\{ E \left[ Q^2 \bar{Z}_k \bar{Z}_k^T \right] - I \right\} \\ & \times \prod_{s=3}^{m+1} \left\{ \hat{E}_s \left[ \dot{P} \dot{B} H_1 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} \left\{ \begin{aligned} & E \left[ Q^2 \bar{Z}_k \bar{Z}_k^T \right] \\ & - \hat{E}_s \left[ Q^2 \bar{Z}_k \bar{Z}_k^T \right] \end{aligned} \right\} \\ & \times \left\{ E \left[ Q^2 \bar{Z}_k \bar{Z}_k^T \right] \right\}^{-1} E \left[ \bar{Z}_k Q^2 \left( \frac{B-\hat{B}}{\hat{B}} \right) \right] \end{aligned} \right\}
\end{aligned}$$

which completes the proof. ■

### Motivation and proofs for Section 4.1.2.

Before proving theorem 33, we provide some calculations as motivation and several preliminary lemmas.

Since  $E_\theta (H_1 B + H_3 | X) = E_\theta (H_1 P + H_2 | X) = 0$ , we can show that

$$\begin{aligned}
& E_\theta \left( \hat{\mathbb{U}}_m \left( (l)_{k(l,0)}^{k(l,1)}, 1 \leq l \leq m-1 \right) \right) \\
& = \left( E_\theta \left( \hat{\epsilon} \bar{Z}_{k(1,0)}^{k(1,1)T} \right) \prod_{u=2}^{m-1} \left[ E_\theta \left( \dot{B} \dot{P} H_1 \bar{Z}_{k(u-1,0)}^{k(u-1,1)} \bar{Z}_{k(u,0)}^{k(u,1)T} - I_{k_{u-1} \times k_u} \right) \right] E_\theta \left( \bar{Z}_{k(m-1,0)}^{k(m-1,1)} \hat{\Delta} \right) \right) \\
& = \left\{ \hat{E} \left( (\delta g + 1) \delta b \bar{Z}_{k(1,0)}^{k(1,1)T} \right) \prod_{u=2}^{m-1} \hat{E} \left( \delta g \hat{Q}^2 \bar{Z}_{k(u-1,0)}^{k(u-1,1)} \bar{Z}_{k(u,0)}^{k(u,1)T} \right) \hat{E} \left( (\delta g + 1) \delta p \bar{Z}_{k(m-1,0)}^{k(m-1,1)} \right) \right\}
\end{aligned}$$

with  $\delta b = \dot{P} \hat{E} (H_1 | X) (\hat{B} - B)$ ,  $\delta p = \dot{B} \hat{E} (H_1 | X) (\hat{P} - P)$ ,  $\delta g = \frac{g-\hat{g}}{\hat{g}}$  and  $\hat{Q}^2 = \dot{B} \dot{P} \hat{E} (H_1 | X)$ .

Below we give a useful representation of  $E_\theta \left[ \widehat{\mathbb{U}}_m \left( (l)_{k(l,0)}^{k(l,1)}, 1 \leq l \leq m-1 \right) \right]$ . Let

$$\begin{aligned}
& B_m \left( \widehat{\mathbb{U}}_m \left( (l)_{k(l,0)}^{k(l,1)}, 1 \leq l \leq m-1 \right) \right) \\
& \equiv \left\{ \widehat{E} \left( \delta b \overline{Z}_{k(1,0)}^{k(1,1)T} \right) \left[ \prod_{u=2}^{m-1} \widehat{E} \left( \delta g \widehat{Q}^2 \overline{Z}_{k(u-1,0)}^{k(u-1,1)} \overline{Z}_{k(u,0)}^{k(u,1)T} \right) \right] \widehat{E} \left( \delta p \overline{Z}_{k(m-1,0)}^{k(m-1,1)} \right) \right\} \\
& B_{m+1}^{bg} \left( \widehat{\mathbb{U}}_m \left( (l)_{k(l,0)}^{k(l,1)}, 1 \leq l \leq m-1 \right) \right) \\
& \equiv \left\{ \widehat{E} \left( \delta g \delta b \overline{Z}_{k(1,0)}^{k(1,1)T} \right) \prod_{u=2}^{m-1} \widehat{E} \left( \delta g \widehat{Q}^2 \overline{Z}_{k(u-1,0)}^{k(u-1,1)} \overline{Z}_{k(u,0)}^{k(u,1)T} \right) \widehat{E} \left( \delta p \overline{Z}_{k(m-1,0)}^{k(m-1,1)} \right) \right\} \\
& B_{m+1}^{pg} \left( \widehat{\mathbb{U}}_m \left( (l)_{k(l,0)}^{k(l,1)}, 1 \leq l \leq m-1 \right) \right) \\
& \equiv \left\{ \widehat{E} \left( \delta b \overline{Z}_{k(1,0)}^{k(1,1)T} \right) \prod_{u=2}^{m-1} \widehat{E} \left( \delta g \widehat{Q}^2 \overline{Z}_{k(u-1,0)}^{k(u-1,1)} \overline{Z}_{k(u,0)}^{k(u,1)T} \right) \widehat{E} \left( \delta g \delta p \overline{Z}_{k(m-1,0)}^{k(m-1,1)} \right) \right\} \\
& B_{m+2} \left( \widehat{\mathbb{U}}_m \left( (l)_{k(l,0)}^{k(l,1)}, 1 \leq l \leq m-1 \right) \right) \\
& \equiv \left\{ \widehat{E} \left( \delta g \delta b \overline{Z}_{k(1,0)}^{k(1,1)T} \right) \prod_{u=2}^{m-1} \widehat{E} \left( \delta g \widehat{Q}^2 \overline{Z}_{k(u-1,0)}^{k(u-1,1)} \overline{Z}_{k(u,0)}^{k(u,1)T} \right) \widehat{E} \left( \delta g \delta p \overline{Z}_{k(m-1,0)}^{k(m-1,1)} \right) \right\}
\end{aligned}$$

Then we may write

$$\begin{aligned}
& E_\theta \left( \widehat{\mathbb{U}}_m \left( (l)_{k(l,0)}^{k(l,1)}, 1 \leq l \leq m-1 \right) \right) \\
& = \left( B_m + B_{m+1}^{bg} + B_{m+1}^{pg} + B_{m+2} \right) \left( \widehat{\mathbb{U}}_m \left( (l)_{k(l,0)}^{k(l,1)}, 1 \leq l \leq m-1 \right) \right)
\end{aligned}$$

We shall require the following:

**Lemma 57** Assume  $k_1 = (k_1(l, s))_{(t-1) \times 2}$  is a  $(t-1) \times 2$  dimensional matrix with

$l \in \{0, 1, \dots, t-2\}$ ,  $s \in \{0, 1\}$ , and  $k_1(0, 0) = k_1(t-2, 0) \equiv 0$ . For  $\forall t > 3$ , if

$$\chi(t, k_1; \theta) \equiv \left\{ \begin{array}{l} \left( \begin{array}{l} B_t \left( \widehat{\mathbb{U}}_{t-2} \left( (l)_{k_1(l,0)}^{k_1(l,1)}, 1 \leq l \leq t-3 \right) \right) \\ -B_t^{bg} \left( \widehat{\mathbb{U}}_{t-1} \left( (l)_{k_1(l,0)}^{k_1(l,1)}, (t-2)_{k_1(t-2,0)}^{k_1(t-2,0)}, 1 \leq l \leq t-3 \right) \right) \end{array} \right) \\ - \left( \begin{array}{l} B_t^{pg} \left( \widehat{\mathbb{U}}_{t-1} \left( (1)_{k_1(0,0)}^{k_1(0,1)}, (l+1)_{k_1(l,0)}^{k_1(l,1)}, 1 \leq l \leq t-3 \right) \right) \\ -B_t \left( \widehat{\mathbb{U}}_t \left( (1)_{k_1(0,0)}^{k_1(0,1)}, (l+1)_{k_1(l,0)}^{k_1(l,1)}, (t-1)_{k_1(t-2,0)}^{k_1(t-2,0)}, 1 \leq l \leq t-3 \right) \right) \end{array} \right) \end{array} \right\}$$

then

$$|\chi(t, k_1; \theta)| = O_p \left( \left( \frac{\log n}{n} \right)^{-\frac{\beta_g}{d+2\beta_g}} (k_1(0, 1))^{-\frac{\beta_b}{d}} (k_1(t-2, 1))^{-\frac{\beta_p}{d}} \right)$$

This lemma explains how to use higher order U-statistics to estimate the  $t$ th order contribution of  $E_\theta \left( \widehat{\mathbb{U}}_{t-2} \left( (l)_{k_1(l,0)}^{k_1(l,1)}, 1 \leq l \leq t-3 \right) \right)$  with a residual bias not exceeding  $\left( \frac{\log n}{n} \right)^{-\frac{\beta_g}{d+2\beta_g}} (k_1(0, 1))^{-\frac{\beta_b}{d}} (k_1(t-2, 1))^{-\frac{\beta_p}{d}}$ . Our estimator uses this idea to reduce fourth and higher order estimation bias to the optimal rate.

**Proof.** (Lemma 57)

$$\begin{aligned}
& \chi(t, k_1; \theta) \\
&= \left\{ \begin{aligned} & \widehat{E} \left( \delta b \delta g \overline{Z}_{k_1(1,0)}^{k_1(1,1)T} \right) \prod_{u=2}^{t-3} \widehat{E} \left( \delta g \widehat{Q}^2 \overline{Z}_{k_1(u-1,0)}^{k_1(u-1,1)} \overline{Z}_{k_1(u,0)}^{k_1(u,1)T} \right) \times \\ & \widehat{E} \left( \delta g \widehat{Q} \overline{Z}_{k_1(t-3,0)}^{k_1(t-3,1)} \left( \widehat{Q}^{-1} \delta p - \widehat{\Pi} \left( \widehat{Q}^{-1} \delta p | \widehat{Q} \overline{Z}_{k_1(t-2,0)}^{k_1(t-2,1)} \right) \right) \right) \end{aligned} \right\} \\
&- \left\{ \begin{aligned} & \widehat{E} \left( \widehat{Q} \delta g \overline{Z}_{k_1(1,0)}^{k_1(1,1)T} \widehat{\Pi} \left( \widehat{Q}^{-1} \delta b | \left( \widehat{Q} \overline{Z}_{k_1(0,0)}^{k_1(0,1)} \right) \right) \right) \prod_{u=2}^{t-3} \widehat{E} \left( \delta g \widehat{Q}^2 \overline{Z}_{k_1(u-1,0)}^{k_1(u-1,1)} \overline{Z}_{k_1(u,0)}^{k_1(u,1)T} \right) \\ & \times \widehat{E} \left( \delta g \widehat{Q} \overline{Z}_{k_1(t-3,0)}^{k_1(t-3,1)} \left( \widehat{Q}^{-1} \delta p - \widehat{\Pi} \left( \widehat{Q}^{-1} \delta p | \left( \widehat{Q} \overline{Z}_{k_1(t-2,0)}^{k_1(t-2,1)} \right) \right) \right) \right) \end{aligned} \right\} \\
&= \left\{ \begin{aligned} & \widehat{E} \left( \widehat{Q} \delta g \overline{Z}_{k_1(1,0)}^{k_1(1,1)T} \widehat{\Pi}^\perp \left( \widehat{Q}^{-1} \delta b | \left( \widehat{Q} \overline{Z}_{k_1(0,0)}^{k_1(0,1)} \right) \right) \right) \prod_{u=2}^{t-3} \widehat{E} \left( \delta g \widehat{Q}^2 \overline{Z}_{k_1(u-1,0)}^{k_1(u-1,1)} \overline{Z}_{k_1(u,0)}^{k_1(u,1)T} \right) \\ & \times \widehat{E} \left( \delta g \widehat{Q} \overline{Z}_{k_1(t-3,0)}^{k_1(t-3,1)} \widehat{\Pi}^\perp \left( \widehat{Q}^{-1} \delta p | \left( \widehat{Q} \overline{Z}_{k_1(t-2,0)}^{k_1(t-2,1)} \right) \right) \right) \end{aligned} \right\} \\
&= \left\{ \widehat{E} \left( \delta g \left[ \left( \prod_{u=1}^{t-3} R_u \right) \left( \widehat{\Pi}^\perp \left( \widehat{Q}^{-1} \delta b | \left( \widehat{Q} \overline{Z}_{k_1(0,0)}^{k_1(0,1)} \right) \right) \right) \right] \widehat{\Pi}^\perp \left( \widehat{Q}^{-1} \delta p | \left( \widehat{Q} \overline{Z}_{k_1(t-2,0)}^{k_1(t-2,1)} \right) \right) \right) \right\}
\end{aligned}$$

where  $R_u(H) = \widehat{E} \left( \delta g \widehat{Q} \overline{Z}_{k_1(u,0)}^{k_1(u,1)T} H \right) \widehat{Q} \overline{Z}_{k_1(u,0)}^{k_1(u,1)} = \widehat{\Pi} \left( \delta g H | \left( \widehat{Q} \overline{Z}_{k_1(u,0)}^{k_1(u,1)} \right) \right)$ ,  $\left( \prod_{u=1}^s R_u \right)(H) = R_s \left[ \left( \prod_{u=1}^{s-1} R_u \right)(H) \right]$ ,  $\widehat{\Pi}(\cdot|\cdot) = \Pi_{\widehat{\theta}}(\cdot|\cdot)$ , and  $\widehat{\Pi}^\perp(H|\Gamma) = H - \widehat{\Pi}(H|\Gamma)$ .

Since projection operator has operator norm of 1, we have

$$\begin{aligned}
& |\chi(t, k_1; \theta)| \\
&\leq \left[ \begin{aligned} & \|\delta g\|_\infty^{t-2} \left\{ \widehat{E} \left( \widehat{\Pi}^\perp \left( \widehat{Q}^{-1} \delta b | \left( \widehat{Q} \overline{Z}_0^{k_1(0,1)} \right) \right) \right)^2 \right\}^{1/2} \\ & \times \left\{ \widehat{E} \left( \widehat{\Pi}^\perp \left( \widehat{Q}^{-1} \delta p | \left( \widehat{Q} \overline{Z}_0^{k_1(t-2,1)} \right) \right) \right)^2 \right\}^{1/2} \end{aligned} \right] \\
&= O_p \left( \left( \frac{\log n}{n} \right)^{-\frac{(t-2)\beta_g}{d+2\beta_g}} (k_1(0,1))^{-\frac{\beta_b}{d}} (k_1(t-2,1))^{-\frac{\beta_p}{d}} \right)
\end{aligned}$$

The last equality holds from theorem (18). ■

Before proving theorem 33, let's first introduce a lemma which will be used in the proof multiple times.

**Lemma 58**  $\forall 2 \times 2$  matrix  $k \equiv (k(l, s))_{2 \times 2}$ ,  $l \in \{1, 2\}$ ,  $s \in \{0, 1\}$ , and  $0 < k(l, 0) < k(l, 1)$ ,

$$\widehat{E} \left[ \left( \widehat{Q} \overline{Z}_{k(l,0)}^{k(l,1)} \right) \left( \widehat{Q} \overline{Z}_{k(l,0)}^{k(l,1)T} \right) \right] = I, \quad l = 1, 2$$

Moreover

$$\begin{aligned} & \widehat{E} \left( \delta b \overline{Z}_{k(1,0)}^{k(1,1)T} \right) \widehat{E} \left( \delta g \widehat{Q}^2 \overline{Z}_{k(1,0)}^{k(1,1)} \overline{Z}_{k(2,0)}^{k(2,1)T} \right) \widehat{E} \left[ \delta p \overline{Z}_{k(2,0)}^{k(2,1)} \right] \\ &= \widehat{E} \left( \delta g \widehat{\Pi} \left( \widehat{Q}^{-1} \delta b \mid \left( \widehat{Q} \overline{Z}_{k(1,0)}^{k(1,1)} \right) \right) \widehat{\Pi} \left( \widehat{Q}^{-1} \delta p \mid \left( \widehat{Q} \overline{Z}_{k(2,0)}^{k(2,1)} \right) \right) \right) \end{aligned}$$

$$\text{and } \widehat{\Pi} \left( H \mid \left( \widehat{Q} \overline{Z}_{k(1,0)}^{k(1,1)} \right) \right) = \widehat{\Pi} \left( H \mid \left( \widehat{Q} \overline{Z}_{k(1,0)}^{c_k} \right) \right) + \widehat{\Pi} \left( H \mid \left( \widehat{Q} \overline{Z}_{c_k}^{k(1,1)} \right) \right)$$

$$\text{if } k(1, 0) < c_k < k(1, 1).$$

**Proof.** The orthonormality of  $\left\{ \widehat{Q} Z_l : l = 1, 2, \dots \right\}$  under  $E_{\widehat{\theta}}$  comes directly from definition.

Hence  $\widehat{\Pi} \left( \widehat{Q}^{-1} \delta b \mid \left( \widehat{Q} \overline{Z}_{k(1,0)}^{k(1,1)} \right) \right) = \widehat{E} \left( \widehat{Q}^{-1} \delta b \widehat{Q} \overline{Z}_{k(1,0)}^{k(1,1)T} \right) \widehat{Q} \overline{Z}_{k(1,0)}^{k(1,1)} = \widehat{E} \left( \delta b \overline{Z}_{k(1,0)}^{k(1,1)T} \right) \widehat{Q} \overline{Z}_{k(1,0)}^{k(1,1)}$ .

And  $\text{lin} \left( \widehat{Q} \overline{Z}_{k(1,0)}^{k(1,1)} \right)$ , the linear space generated by  $\widehat{Q} \overline{Z}_{k(1,0)}^{k(1,1)}$ , equals  $\left( \widehat{Q} \overline{Z}_{k(1,0)}^{c_k} \right) \oplus \left( \widehat{Q} \overline{Z}_{c_k}^{k(1,1)} \right)$ . ■

**Proof.** (Theorem 33) Let  $\mathbb{H}_v^* = \mathbb{G}(s, v) = \mathbb{Q}_v \equiv 0$  for  $v > m(\beta_g, \beta_b, \beta_p)$ , and

$\mathbb{G}(s, 2) = \mathbb{Q}_2 \equiv 0$ . For  $\forall 3 < t$ , define

$$\begin{aligned}
B_t^{(H^*)} &= (-1)^{t-1} \left( B_t(\mathbb{H}_{t-2}^*) - B_t^{bg}(\mathbb{H}_{t-1}^*) - B_t^{pg}(\mathbb{H}_{t-1}^*) + B_t(\mathbb{H}_t^*) \right) \\
B_t^{(G)} &= \sum_{s=1}^J B_t^{(G,s)} \\
B_t^{(G,s)} &= (-1)^{t-1} \begin{pmatrix} B_t(\mathbb{G}(s, t-2)) - B_t^{bg}(\mathbb{G}(s, t-1)) \\ -B_t^{pg}(\mathbb{G}(s, t-1)) + B_t(\mathbb{G}(s, t)) \end{pmatrix} \\
B_t^{(Q)} &= (-1)^{t-1} \left( B_t(\mathbb{Q}_{t-2}) - B_t^{bg}(\mathbb{Q}_{t-1}) - B_t^{pg}(\mathbb{Q}_{t-1}) + B_t(\mathbb{Q}_t) \right)
\end{aligned}$$

then

$$\begin{aligned}
&E \left( \widehat{\psi}_{\mathcal{K}_J}^{eff}(\beta_g, \beta_b, \beta_p) \right) - \psi(\theta) \\
&= \left[ E_\theta \left( H_1 \widehat{B} \widehat{P} + H_2 \widehat{B} + H_3 \widehat{P} + H_4 \right) - \psi(\theta) \right]_{-(L4.1)} \\
&- B_2(\mathbb{H}_2^*) + B_3(\mathbb{H}_3^* - \mathbb{H}_2^*) + \sum_{s=1}^J B_3(\mathbb{G}(s, 3)) + B_3(\mathbb{Q}_3) \\
&+ \sum_{t=4}^{m(\beta_g, \beta_b, \beta_p)+2} \left( B_t^{(H^*)} + B_t^{(G)} + B_t^{(Q)} \right)
\end{aligned}$$

$$\begin{aligned}
(L4.1) &= E_\theta \left( H_1 \left( B - \widehat{B} \right) \left( P - \widehat{P} \right) \right) \\
&= \widehat{E} \left[ \delta g \widehat{E}(H_1|X) \left( B - \widehat{B} \right) \left( P - \widehat{P} \right) \right] \\
&+ \widehat{E} \left( \widehat{E}(H_1|X) \left( B - \widehat{B} \right) \left( P - \widehat{P} \right) \right) \\
&= \widehat{E} \left( \delta g \widehat{Q}^{-2} \delta b \delta p \right)_{-(L4.2)} + \widehat{E} \left( \widehat{Q}^{-2} \delta b \delta p \right)_{-(L4.3)}
\end{aligned}$$

$$\begin{aligned}
& (L4.3) - B_2(H_2^*) \\
&= \widehat{E} \left( \widehat{Q}^{-2} \delta b \delta p \right) - \widehat{E} \left( \delta b \overline{Z}_0^{k-1T} \right) \widehat{E} \left( \delta p \overline{Z}_0^{k-1} \right) \\
&= \widehat{E} \left( \widehat{Q}^{-2} \delta b \delta p \right) - \widehat{E} \left[ \widehat{\Pi} \left( \widehat{Q}^{-1} \delta b | \widehat{Q} \overline{Z}_0^{k-1} \right) \widehat{\Pi} \left( \widehat{Q}^{-1} \delta p | \widehat{Q} \overline{Z}_0^{k-1} \right) \right] \\
&= \widehat{E} \left[ \widehat{\Pi}^\perp \left( \widehat{Q}^{-1} \delta b | \left( \widehat{Q} \overline{Z}_0^{k-1} \right) \right) \widehat{\Pi}^\perp \left( \widehat{Q}^{-1} \delta p | \left( \widehat{Q} \overline{Z}_0^{k-1} \right) \right) \right]
\end{aligned}$$

Then

$$\begin{aligned}
& |(L4.3) - B_2(H_2^*)| = O_p \left( k_{-1}^{2\beta/d} \right) \\
& (L4.2) - B_3(\mathbb{H}_2^*) + B_3(\mathbb{H}_2^*) + \sum_{s=1}^J B_3(\mathbb{G}(s, 3)) + B_3(\mathbb{Q}_3) \\
&= \left\{ \begin{aligned} & \widehat{E} \left( \widehat{Q}^{-2} \delta b \delta p \delta g \right) - \widehat{E} \left( \delta b \delta g \overline{Z}_0^{k-1T} \right) \widehat{E} \left( \delta p \overline{Z}_0^{k-1} \right) \\ & - \widehat{E} \left( \delta b \overline{Z}_0^{k-1T} \right) \widehat{E} \left( \delta p \delta g \overline{Z}_0^{k-1} \right) \end{aligned} \right\}_{-(L5.1)} \\
&+ \widehat{E} \left( \delta b \overline{Z}_0^{k_0T} \right) \widehat{E} \left( \delta g \widehat{Q}^2 \overline{Z}_0^{k_0} \overline{Z}_0^{k_0T} \right) \widehat{E} \left[ \delta p \overline{Z}_0^{k_0} \right] \\
&+ \widehat{E} \left( \delta b \overline{Z}_{k_0}^{k-1T} \right) \widehat{E} \left( \delta g \widehat{Q}^2 \overline{Z}_{k_0}^{k-1} \overline{Z}_0^{k_0T} \right) \widehat{E} \left[ \delta p \overline{Z}_0^{k_0} \right] \\
&+ \widehat{E} \left( \delta b \overline{Z}_0^{k_0T} \right) \widehat{E} \left( \delta g \widehat{Q}^2 \overline{Z}_0^{k_0} \overline{Z}_{k_0}^{k-1T} \right) \widehat{E} \left[ \delta p \overline{Z}_{k_0}^{k-1} \right] \\
&+ \sum_{s=1}^J \left\{ \begin{aligned} & \widehat{E} \left( \delta b \overline{Z}_{k_{2s-2}}^{k_{2s-1}T} \right) \widehat{E} \left( \delta g \widehat{Q}^2 \overline{Z}_{k_{2s-2}}^{k_{2s-1}} \overline{Z}_{k_{2s-2}}^{k_{2s}T} \right) \widehat{E} \left[ \delta p \overline{Z}_{k_{2s-2}}^{k_{2s}} \right] \\ & + \widehat{E} \left( \delta b \overline{Z}_{k_{2s-2}}^{k_{2s}T} \right) \widehat{E} \left( \delta g \widehat{Q}^2 \overline{Z}_{k_{2s-2}}^{k_{2s}} \overline{Z}_{k_{2s}}^{k_{2s-1}T} \right) \widehat{E} \left[ \delta p \overline{Z}_{k_{2s}}^{k_{2s-1}} \right] \end{aligned} \right\} \\
&+ \widehat{E} \left( \delta b \overline{Z}_{k_{2J}}^{k_{2J+1}T} \right) \widehat{E} \left( \delta g \widehat{Q}^2 \overline{Z}_{k_{2J}}^{k_{2J+1}} \overline{Z}_{k_{2J}}^{k_{2J+1}T} \right) \widehat{E} \left[ \delta p \overline{Z}_{k_{2J}}^{k_{2J+1}} \right]
\end{aligned}$$

because

$$\begin{aligned}
& (L5.1) \\
& = \widehat{E} \left( \begin{aligned} & \delta g \left\{ \widehat{\Pi} \left( \widehat{Q}^{-1} \delta b \mid \left( \widehat{Q} \overline{Z}_0^{k-1} \right) \right) + \widehat{\Pi}^\perp \left( \widehat{Q}^{-1} \delta b \mid \left( \widehat{Q} \overline{Z}_0^{k-1} \right) \right) \right\} \\ & \times \left\{ \widehat{\Pi} \left( \widehat{Q}^{-1} \delta p \mid \left( \widehat{Q} \overline{Z}_0^{k-1} \right) \right) + \widehat{\Pi}^\perp \left( \widehat{Q}^{-1} \delta p \mid \left( \widehat{Q} \overline{Z}_0^{k-1} \right) \right) \right\} \end{aligned} \right) \\
& - \widehat{E} \left( \delta g \delta p \widehat{Q}^{-1} \widehat{\Pi} \left( \widehat{Q}^{-1} \delta b \mid \left( \widehat{Q} \overline{Z}_0^{k-1} \right) \right) \right) \\
& - \widehat{E} \left( \delta g \delta b \widehat{Q}^{-1} \widehat{\Pi} \left( \widehat{Q}^{-1} \delta p \mid \left( \widehat{Q} \overline{Z}_0^{k-1} \right) \right) \right) \\
& = \widehat{E} \left( \delta g \widehat{\Pi}^\perp \left( \widehat{Q}^{-1} \delta b \mid \left( \widehat{Q} \overline{Z}_0^{k-1} \right) \right) \widehat{\Pi}^\perp \left( \widehat{Q}^{-1} \delta p \mid \left( \widehat{Q} \overline{Z}_0^{k-1} \right) \right) \right) \\
& - \widehat{E} \left( \delta g \widehat{\Pi} \left( \widehat{Q}^{-1} \delta b \mid \left( \widehat{Q} \overline{Z}_0^{k-1} \right) \right) \widehat{\Pi} \left( \widehat{Q}^{-1} \delta p \mid \left( \widehat{Q} \overline{Z}_0^{k-1} \right) \right) \right)
\end{aligned}$$

and by lemma 58, we now have

$$\begin{aligned}
& (L4.2) - B_3(\mathbb{H}_2^*) + B_3(\mathbb{H}_2^*) + \sum_{s=1}^J B_3(\mathbb{G}(s, 3)) + B_3(\mathbb{Q}_3) \\
& = \widehat{E} \left( \delta g \widehat{\Pi}^\perp \left( \widehat{Q}^{-1} \delta b \mid \left( \widehat{Q} \overline{Z}_0^{k-1} \right) \right) \widehat{\Pi}^\perp \left( \widehat{Q}^{-1} \delta p \mid \left( \widehat{Q} \overline{Z}_0^{k-1} \right) \right) \right)_{-(L5.2)} \\
& - \sum_{s=0}^J \widehat{E} \left[ \delta g \widehat{\Pi} \left( \widehat{Q}^{-1} \delta b \mid \left( \widehat{Q} \overline{Z}_{k_{2s}}^{k_{2s}-1} \right) \right) \widehat{\Pi} \left( \widehat{Q}^{-1} \delta p \mid \left( \widehat{Q} \overline{Z}_{k_{2s}+1}^{k_{2s}-1} \right) \right) \right]_{-(L5.3)} \\
& + \widehat{E} \left[ \delta g \widehat{\Pi} \left( \widehat{Q}^{-1} \delta b \mid \left( \widehat{Q} \overline{Z}_{k_{2s}+1}^{k_{2s}-1} \right) \right) \widehat{\Pi} \left( \widehat{Q}^{-1} \delta p \mid \left( \widehat{Q} \overline{Z}_{k_{2s}}^{k_{2s}+1} \right) \right) \right]_{-(L5.4)}
\end{aligned}$$



Therefore it is easy to show from theorem 18 that

$$\begin{aligned}
|(L5.2)| &= O_p \left( \left( \frac{\log n}{n} \right)^{-\frac{\beta_g}{d+2\beta_g}} k_{-1}^{2\beta/d} \right) \\
|(L5.3)| &= O_p \left( \left( \frac{\log n}{n} \right)^{-\frac{\beta_g}{d+2\beta_g}} k_{2s}^{-\beta_b/d} k_{2s+1}^{-\beta_p/d} \right) \\
|(L5.4)| &= O_p \left( \left( \frac{\log n}{n} \right)^{-\frac{\beta_g}{d+2\beta_g}} k_{2s+1}^{-\beta_b/d} k_{2s}^{-\beta_p/d} \right)
\end{aligned}$$

$$\forall 3 < t \leq m(\beta_g, \beta_b, \beta_p),$$

$$\begin{aligned}
&(-1)^{t-1} B_t^{(H^*)} \\
&= \left( B_t(\mathbb{H}_{t-2}^*) - B_t^{bg}(\mathbb{H}_{t-1}^*) - B_t^{pg}(\mathbb{H}_{t-1}^*) + B_t(\mathbb{H}_t^*) \right) \\
&= \left\{ \begin{array}{l} B_t \left( \widehat{\mathbb{U}}_{t-2} \left( \begin{smallmatrix} k_0 \\ 0 \end{smallmatrix} \right) \right) - B_t^{bg} \left( \widehat{\mathbb{U}}_{t-1} \left( \begin{smallmatrix} k_0 \\ 0 \end{smallmatrix} \right) \right) \\ - \left[ B_t^{pg} \left( \widehat{\mathbb{U}}_{t-1} \left( \begin{smallmatrix} k_0 \\ 0 \end{smallmatrix} \right) \right) - B_t \left( \widehat{\mathbb{U}}_t \left( \begin{smallmatrix} k_0 \\ 0 \end{smallmatrix} \right) \right) \right] \end{array} \right\}_{-(L6.1)} \\
&+ \sum_{u=1}^{t-3} \left\{ \begin{array}{l} \left[ B_t \left( \widehat{\mathbb{U}}_{t-2}^{(u)} \left( \begin{smallmatrix} k_{-1} & k_0 \\ k_0 & 0 \end{smallmatrix} \right) \right) - B_t^{bg} \left( \widehat{\mathbb{U}}_{t-1}^{(u)} \left( \begin{smallmatrix} k_{-1} & k_0 \\ k_0 & 0 \end{smallmatrix} \right) \right) \right] \\ - \left[ B_t^{pg} \left( \widehat{\mathbb{U}}_{t-1}^{(u+1)} \left( \begin{smallmatrix} k_{-1} & k_0 \\ k_0 & 0 \end{smallmatrix} \right) \right) - B_t \left( \widehat{\mathbb{U}}_t^{(u+1)} \left( \begin{smallmatrix} k_{-1} & k_0 \\ k_0 & 0 \end{smallmatrix} \right) \right) \right] \end{array} \right\}_{-(L6.2.u)} \\
&- B_t^{pg} \left( \widehat{\mathbb{U}}_{t-1}^{(1)} \left( \begin{smallmatrix} k_{-1} & k_0 \\ k_0 & 0 \end{smallmatrix} \right) \right) - B_t \left( \widehat{\mathbb{U}}_t^{(1)} \left( \begin{smallmatrix} k_{-1} & k_0 \\ k_0 & 0 \end{smallmatrix} \right) \right)_{-(L6.3)} \\
&- B_t^{bg} \left( \widehat{\mathbb{U}}_{t-1}^{(t-2)} \left( \begin{smallmatrix} k_{-1} & k_0 \\ k_0 & 0 \end{smallmatrix} \right) \right) - B_t \left( \widehat{\mathbb{U}}_t^{(t-1)} \left( \begin{smallmatrix} k_{-1} & k_0 \\ k_0 & 0 \end{smallmatrix} \right) \right)_{-(L6.4)}
\end{aligned}$$

It can be shown that  $(L6.1) = \chi(t, k_1; \theta)$  with  $k_1(l, 0) \equiv 0$ ,  $k_1(l, 1) \equiv k_0$ ,  $0 \leq l \leq t-2$ ;

$(L6.2.u) = \chi(t, k_1; \theta)$  with  $k_1(l, 0) = 0$ ,  $k_1(l, 1) = k_0$  for  $l \neq u$  and  $k_1(u, 0) = k_0$ ,

$k_1(u, 1) = k_{-1}$ . And

$$\begin{aligned}
& |(L6.3)| \\
&= \left| \left\{ \begin{aligned} & \widehat{E} \left( \delta g \widehat{Q} \widehat{\Pi} \left( \widehat{Q}^{-1} \delta b \mid \left( \widehat{Q} \overline{Z}_{k_0}^{k_{-1}} \right) \right) \overline{Z}_0^{k_0 T} \right) \left( \widehat{E} \left( \delta g \widehat{Q}^2 \overline{Z}_0^{k_0} \overline{Z}_0^{k_0 T} \right) \right)^{t-4} \\ & \times E \left( \delta g \widehat{Q} \overline{Z}_0^{k_0} \widehat{\Pi}^\perp \left( \widehat{Q}^{-1} \delta b \mid \left( \widehat{Q} \overline{Z}_0^{k_0} \right) \right) \right) \end{aligned} \right\} \right| \\
&= O_p \left( \left( \frac{\log n}{n} \right)^{-\frac{(t-2)\beta_g}{d+2\beta_g}} k_0^{-2\beta/d} \right)
\end{aligned}$$

Similarly

$$|(L6.4)| = O_p \left( \left( \frac{\log n}{n} \right)^{-\frac{(t-2)\beta_g}{d+2\beta_g}} k_0^{-2\beta/d} \right)$$

From lemma 57, we now have  $|(L6.1)| = O_p \left( \left( \frac{\log n}{n} \right)^{-\frac{(t-2)\beta_g}{d+2\beta_g}} k_0^{-2\beta/d} \right)$  therefore

$$\left| B_t^{(H^*)} \right| = O_p \left( \left( \frac{\log n}{n} \right)^{-\frac{(t-2)\beta_g}{d+2\beta_g}} k_0^{-2\beta/d} \right)$$

$\forall t \geq 5$ , following the same argument as above, we know as well

$$\begin{aligned}
\left| B_t^{(G)} \right| &= O_p \left( \left( \frac{\log n}{n} \right)^{-\frac{(t-2)\beta_g}{d+2\beta_g}} k_0^{-2\beta/d} \right) \\
\left| B_t^{(Q)} \right| &= O_p \left( \left( \frac{\log n}{n} \right)^{-\frac{(t-2)\beta_g}{d+2\beta_g}} k_0^{-2\beta/d} \right)
\end{aligned}$$

$\forall 1 \leq s \leq J$

$$\begin{aligned}
B_4^{(G,s)} &= B_4 \left( \widehat{\mathbb{U}}_3^{(1,2)} \left( \begin{matrix} k_{2s-1} & k_{2s} & k_0 \\ k_{2s-2} & k_{2s-2} & 0 \end{matrix} \right) \right) + B_4 \left( \widehat{\mathbb{U}}_3^{(1,2)} \left( \begin{matrix} k_{2s} & k_{2s-1} & k_0 \\ k_{2s-2} & k_{2s} & 0 \end{matrix} \right) \right) \\
&- B_4 \left( \widehat{\mathbb{U}}_4^{(1,2)} \left( \begin{matrix} k_{2s-1} & k_{2s} & k_0 \\ k_{2s-2} & k_{2s-2} & 0 \end{matrix} \right) \right) + \widehat{\mathbb{U}}_4^{(2,3)} \left( \begin{matrix} k_{2s-1} & k_{2s} & k_0 \\ k_{2s-2} & k_{2s-2} & 0 \end{matrix} \right) \\
&- B_4 \left( \widehat{\mathbb{U}}_4^{(1,2)} \left( \begin{matrix} k_{2s} & k_{2s-1} & k_0 \\ k_{2s-2} & k_{2s} & 0 \end{matrix} \right) \right) + \widehat{\mathbb{U}}_4^{(2,3)} \left( \begin{matrix} k_{2s} & k_{2s-1} & k_0 \\ k_{2s-2} & k_{2s} & 0 \end{matrix} \right)
\end{aligned}$$

From lemma 58

$$\begin{aligned}
& B_4 \left( \widehat{\mathbb{U}}_3^{(1,2)} \begin{pmatrix} k_{2s-1} & k_{2s} & k_0 \\ k_{2s-2} & k_{2s-2} & 0 \end{pmatrix} \right) - B_4 \left( \widehat{\mathbb{U}}_4^{(1,2)} \begin{pmatrix} k_{2s-1} & k_{2s} & k_0 \\ k_{2s-2} & k_{2s-2} & 0 \end{pmatrix} + \widehat{\mathbb{U}}_4^{(2,3)} \begin{pmatrix} k_{2s-1} & k_{2s} & k_0 \\ k_{2s-2} & k_{2s-2} & 0 \end{pmatrix} \right) \\
&= \widehat{E} \left( \delta b \delta g \overline{Z}_{k_{2s-2}}^{k_{2s-1}T} \right) \widehat{E} \left( \delta g \widehat{Q}^2 \overline{Z}_{k_{2s-2}}^{k_{2s-1}} \overline{Z}_{k_{2s-2}}^{k_{2s}T} \right) \widehat{E} \left( \delta p \overline{Z}_{k_{2s-2}}^{k_{2s}} \right) \\
&- \widehat{E} \left( \delta b \overline{Z}_0^{k_0T} \right) \widehat{E} \left( \delta g \widehat{Q}^2 \overline{Z}_0^{k_0} \overline{Z}_{k_{2s-2}}^{k_{2s-1}T} \right) \widehat{E} \left( \delta g \widehat{Q}^2 \overline{Z}_{k_{2s-2}}^{k_{2s-1}} \overline{Z}_{k_{2s-2}}^{k_{2s}T} \right) \widehat{E} \left[ \delta p \overline{Z}_{k_{2s-2}}^{k_{2s}} \right] \\
&+ \widehat{E} \left( \delta b \overline{Z}_{k_{2s-2}}^{k_{2s-1}T} \right) \widehat{E} \left( \delta g \widehat{Q}^2 \overline{Z}_{k_{2s-2}}^{k_{2s-1}} \overline{Z}_{k_{2s-2}}^{k_{2s}T} \right) \widehat{E} \left( \delta p \delta g \overline{Z}_{k_{2s-2}}^{k_{2s}} \right) \\
&- \widehat{E} \left( \delta b \overline{Z}_{k_{2s-2}}^{k_{2s-1}T} \right) \widehat{E} \left( \delta g \widehat{Q}^2 \overline{Z}_{k_{2s-2}}^{k_{2s-1}} \overline{Z}_{k_{2s-2}}^{k_{2s}T} \right) \widehat{E} \left( \delta g \widehat{Q}^2 \overline{Z}_{k_{2s-2}}^{k_{2s}} \overline{Z}_0^{k_0T} \right) \widehat{E} \left[ \delta p \overline{Z}_0^{k_0} \right] \\
&= \widehat{E} \left( \widehat{\Pi}^\perp \left( \widehat{Q}^{-1} \delta b | \left( \widehat{Q} \overline{Z}_0^{k_0} \right) \right) \delta g \overline{Z}_{k_{2s-2}}^{k_{2s-1}T} \right) \widehat{E} \left( \delta g \widehat{Q} \overline{Z}_{k_{2s-2}}^{k_{2s-1}} \widehat{\Pi} \left( \widehat{Q}^{-1} \delta p | \left( \widehat{Q} \overline{Z}_{k_{2s-2}}^{k_{2s}} \right) \right) \right) \\
&+ \widehat{E} \left( \widehat{\Pi} \left( \widehat{Q}^{-1} \delta b | \left( \widehat{Q} \overline{Z}_{k_{2s-2}}^{k_{2s-1}} \right) \right) \delta g \overline{Z}_{k_{2s-2}}^{k_{2s}T} \right) \widehat{E} \left( \delta g \widehat{Q} \overline{Z}_{k_{2s-2}}^{k_{2s}} \widehat{\Pi}^\perp \left( \widehat{Q}^{-1} \delta p | \left( \widehat{Q} \overline{Z}_0^{k_0} \right) \right) \right)
\end{aligned}$$

Similarly,

$$\begin{aligned}
& B_4 \left( \widehat{\mathbb{U}}_3^{(1,2)} \begin{pmatrix} k_{2s} & k_{2s-1} & k_0 \\ k_{2s-2} & k_{2s} & 0 \end{pmatrix} \right) \\
&- B_4 \left( \widehat{\mathbb{U}}_4^{(1,2)} \begin{pmatrix} k_{2s} & k_{2s-1} & k_0 \\ k_{2s-2} & k_{2s} & 0 \end{pmatrix} + \widehat{\mathbb{U}}_4^{(2,3)} \begin{pmatrix} k_{2s} & k_{2s-1} & k_0 \\ k_{2s-2} & k_{2s} & 0 \end{pmatrix} \right) \\
&= \widehat{E} \left( \widehat{\Pi}^\perp \left( \widehat{Q}^{-1} \delta b | \left( \widehat{Q} \overline{Z}_0^{k_0} \right) \right) \delta g \overline{Z}_{k_{2s-2}}^{k_{2s}T} \right) \widehat{E} \left( \delta g \widehat{Q} \overline{Z}_{k_{2s-2}}^{k_{2s}} \widehat{\Pi} \left( \widehat{Q}^{-1} \delta p | \left( \widehat{Q} \overline{Z}_{k_{2s}}^{k_{2s-1}} \right) \right) \right) \\
&+ \widehat{E} \left( \widehat{\Pi} \left( \widehat{Q}^{-1} \delta b | \left( \widehat{Q} \overline{Z}_{k_{2s-2}}^{k_{2s}} \right) \right) \delta g \overline{Z}_{k_{2s}}^{k_{2s-1}T} \right) \widehat{E} \left( \delta g \widehat{Q} \overline{Z}_{k_{2s}}^{k_{2s-1}} \widehat{\Pi}^\perp \left( \widehat{Q}^{-1} \delta p | \left( \widehat{Q} \overline{Z}_0^{k_0} \right) \right) \right)
\end{aligned}$$

Therefore

$$\left| B_4^{(G,s)} \right| = O_p \left( \left( \frac{\log n}{n} \right)^{\frac{2\beta_g}{d+2\beta_g}} \max \left( k_{2s-2}^{-\beta_b/d} k_0^{-\beta_p/d}, k_0^{-\beta_b/d} k_{2s-2}^{-\beta_p/d} \right) \right)$$

Applying the above argument to  $B_4^{(Q)}$ , we can show that:

$$\left| B_4^{(Q)} \right| = O_p \left( \left( \frac{\log n}{n} \right)^{\frac{2\beta_g}{d+2\beta_g}} \max \left( k_{2J}^{-\beta_b/d} k_0^{-\beta_p/d}, k_0^{-\beta_b/d} k_{2J}^{-\beta_p/d} \right) \right)$$

In addition,  $\forall L \in \{H^*, G, Q\}$ ,  $t > m(\beta_g, \beta_b, \beta_p)$ ,

$$\left| B_t^{(L)} \right| = O_p \left( \|\delta g\|_\infty^{(t-2)} \|\delta b\|_2 \|\delta p\|_2 \right)$$

which completes the proof of bias. The order of variance follows directly from theorem 26. ■

**Proof.** (Theorem 42 (iii)) As proved in (ii),

$$\psi_\tau i f_{1,\tau(\cdot)}(o; \theta) = -i f_{1,\psi(\tau^\dagger, \cdot)}(o; \tau, \theta). \quad (72)$$

By part 5c) of Theorem 3, consider any suitably smooth one dimensional parametric submodel  $\tilde{\theta}(\zeta)$  with range containing  $\theta$  and contained in  $\Theta(\tau^\dagger)$ , and differentiate both sides of eq. (72) wrt.  $\zeta$ . Then,

$$\begin{aligned} & \psi_\tau \frac{\partial i f_{1,\tau(\cdot)}(o; \theta)}{\partial \zeta} + \left( \psi_{\tau\tau} \frac{\partial \tau}{\partial \zeta} + \frac{\partial \psi_\tau(\tau^\dagger, \tilde{\theta}(\zeta))}{\partial \zeta} \Big|_{\tilde{\theta}(\zeta)=\theta} \right) i f_{1,\tau(\cdot)}(o; \theta) \\ &= - \frac{\partial i f_{1,\psi(\tau^\dagger, \cdot)}(o; \tau, \theta)}{\partial \tau} \Big|_{\tau=\tau^\dagger} \frac{\partial \tau}{\partial \zeta} - \frac{\partial i f_{1,\psi(\tau^\dagger, \cdot)}(o; \tau^\dagger, \theta)}{\partial \zeta} \\ &= - \frac{\partial i f_{1,\psi(\tau^\dagger, \cdot)}(o; \theta)}{\partial \tau} E_\theta [i f_{1,\tau(\cdot)}(O_2; \theta) S_\zeta(\theta)] \\ &= - E_\theta [i f_{1,i f_{1,\psi}(o; \cdot)}(O_2; \theta) S_\zeta(\theta)]. \end{aligned}$$

Further,  $\frac{\partial \tau}{\partial \zeta} = E_\theta [i f_{1,\tau(\cdot)}(O_2, \theta) S_\zeta(\theta)]$  and

$$\begin{aligned} \frac{\partial \psi_\tau(\tau^\dagger, \tilde{\theta}(\zeta))}{\partial \zeta} \Big|_{\tilde{\theta}(\zeta)=\theta} &= \frac{\partial E_\theta [i f_{1,\psi(\tau, \cdot)}(O_2; \theta) S_\zeta(\theta)]}{\partial \tau} \Big|_{\tau=\tau^*} \\ &= E_\theta \left[ \frac{\partial i f_{1,\psi(\tau^\dagger, \cdot)}(O_2; \theta)}{\partial \tau} S_\zeta(\theta) \right] \end{aligned}$$

Thus,

$$\begin{aligned}
& \psi_\tau \frac{\partial i f_{1,\tau(\cdot)}(o; \theta)}{\partial \zeta} \\
&= - \frac{\partial i f_{1,\psi(\tau^\dagger, \cdot)}(o; \theta)}{\partial \tau} E_\theta [i f_{1,\tau(\cdot)}(O_2; \theta) S_\zeta(\theta)] \\
&\quad - E_\theta [i f_{1,if_{1,\psi(o; \cdot)}}(O_2, \theta) S_\zeta(\theta)] \\
&\quad - i f_{1,\tau(\cdot)}(o; \theta) \left( \begin{aligned} & \psi_{\tau\tau} E_\theta [i f_{1,\tau(\cdot)}(O_2, \theta) S_\zeta(\theta)] \\ & + E_\theta \left[ \frac{\partial i f_{1,\psi(\tau^\dagger, \cdot)}(O_2; \theta)}{\partial \tau} S_\zeta(\theta) \right] \end{aligned} \right)
\end{aligned}$$

for any  $o \in \mathcal{O}$  wp. 1. That is, there exists a first order influence function for  $i f_{1,\tau(\cdot)}(o; \theta)$ , and

$$\begin{aligned}
& \mathbb{IF}_{2,2,\tau(\cdot)}(\theta) \\
&= \frac{1}{2} \mathbb{V} \left\{ d_{2,\theta} [i f_{1,if_{1,\tau(\cdot)}(O_1; \cdot)}(O_2; \theta)] \right\} \\
&= -\psi_\tau^{-1} \left\{ \begin{aligned} & \mathbb{IF}_{2,2,\tau(\cdot)}(\theta) + \frac{1}{2} \psi_{\tau\tau} \mathbb{V} (i f_{1,\tau(\cdot)}(O_1; \theta) i f_{1,\tau(\cdot)}(O_2, \theta)) \\ & + \frac{1}{2} \mathbb{V} \left( \begin{aligned} & i f_{1,\tau(\cdot)}(O_1; \theta) d_{1,\theta} \left( \frac{\partial i f_{1,\psi(\tau^\dagger, \cdot)}(O_2; \theta)}{\partial \tau} \right) \\ & + d_{1,\theta} \left( \frac{\partial i f_{1,\psi(\tau^\dagger, \cdot)}(O_1; \theta)}{\partial \tau} \right) i f_{1,\tau(\cdot)}(O_2; \theta) \end{aligned} \right) \end{aligned} \right\}
\end{aligned}$$

which completes the proof. Note that  $d_{m,\theta}(\cdot)$  is defined in eq. (4). ■

**Proof.** (Part ii and iii of Theorem 43)

(iii) The formulae for  $\mathbb{IF}_{1,\tau(\cdot)}(\hat{\theta}) = \mathbb{IF}_{1,\tilde{\tau}_k(\cdot)}(\hat{\theta})$  and  $\mathbb{IF}_{2,2,\tilde{\tau}_k(\cdot)}(\hat{\theta})$  follow from Theorem 42. To verify the formula for  $Q_{2,2,\tilde{\tau}_k(\cdot)}(\hat{\theta})$ , substitute  $\psi_{\tau^2}(\tau^\dagger, \hat{\theta}) = 0$  and  $\partial IF_{1,\tilde{\psi}_k(\tau, \cdot), i_1}(\hat{\theta}) / \partial \tau = -\{A - \hat{p}(X)\}_{i_1}^2$  in the formula for  $Q_{2,2,\tilde{\tau}_k(\cdot)}(\hat{\theta})$  in Theorem 42.

(ii): To obtain Eq (53), note by Theorem 38, we have

$$\begin{aligned}
& var_{\hat{\theta}} \left\{ \mathbb{ES}_{2, \tilde{\tau}_k(\cdot)}^{test} \left( \hat{\theta}(\tau^\dagger) \right) \right\}^{-1} \mathbb{ES}_{2, \tilde{\tau}_k(\cdot)}^{test} \left( \hat{\theta}(\tau^\dagger) \right) \\
&= E_{\hat{\theta}(\tau^\dagger)} \left[ \mathbb{IF}_{2, \tilde{\psi}_k(\tau^\dagger, \cdot)} \left( \hat{\theta}(\tau^\dagger) \right) \mathbb{ES}_{1, \tilde{\tau}_k(\cdot)}^{test} \left( \hat{\theta}(\tau^\dagger) \right) \right]^{-1} \times \\
&\Pi_{\hat{\theta}(\tau^\dagger)} \left[ \mathbb{IF}_{2, \tilde{\psi}_k(\tau^\dagger, \cdot)} \left( \hat{\theta}(\tau^\dagger) \right) | \Gamma_2^{test} \left( \hat{\theta}(\tau^\dagger), \tau^\dagger \right) \right]
\end{aligned}$$

But by Theorem 42 and the definition of  $\mathbb{ES}_1^{test}$ , we have

$$\begin{aligned}
&\mathbb{ES}_{1, \tilde{\tau}_k(\cdot)}^{test} \left( \hat{\theta}(\tau^\dagger) \right) = \mathbb{ES}_{1, \tau(\cdot)}^{test} \left( \hat{\theta}(\tau^\dagger) \right) \\
&= v \left( \hat{\theta}(\tau^\dagger) \right) E_{\hat{\theta}(\tau^\dagger)} \left[ \left\{ Y^*(\tau^\dagger) - \hat{b}(X, \tau^\dagger) \right\}^2 \{A - \hat{p}(X)\}^2 \right]^{-1} \\
&\quad \times \left\{ Y^*(\tau^\dagger) - \hat{b}(X, \tau^\dagger) \right\} \{A - \hat{p}(X)\}
\end{aligned}$$

thus, we obtain  $E_{\hat{\theta}(\tau^\dagger)} \left[ \mathbb{IF}_{2, \tilde{\psi}_k(\tau^\dagger, \cdot)} \left( \hat{\theta}(\tau^\dagger) \right) \mathbb{ES}_{1, \tilde{\tau}_k(\cdot)}^{test} \left( \hat{\theta}(\tau^\dagger) \right) \right]^{-1} = v \left( \hat{\theta}(\tau^\dagger) \right)^{-1}$

Now

$$\begin{aligned}
&\Pi_{\hat{\theta}} \left[ \mathbb{IF}_{2, \tilde{\psi}_k(\tau^\dagger, \cdot)} \left( \hat{\theta}(\tau^\dagger) \right) | \Gamma_2^{test} \left( \hat{\theta}(\tau^\dagger), \tau^\dagger \right) \right] \\
&= \mathbb{IF}_{2, \tilde{\psi}_k(\tau^\dagger, \cdot)} \left( \hat{\theta}(\tau^\dagger) \right) - \Pi_{\hat{\theta}} \left[ \mathbb{IF}_{2, \tilde{\psi}_k(\tau^\dagger, \cdot)} \left( \hat{\theta}(\tau^\dagger) \right) | \left\{ \mathbb{U}_{2, 2, \tilde{\tau}_k(\cdot)}^{test, \perp} \left( \hat{\theta}(\tau^\dagger), \tau^\dagger \right) \right\} \right]
\end{aligned}$$

Let  $\hat{\epsilon}$  denote  $Y - \hat{b}(X)$ , and  $\hat{\Delta}$  denote  $A - \hat{p}(X)$ . Next, we show that

$$\Pi_{\hat{\theta}} \left[ \mathbb{IF}_{2, \tilde{\psi}_k(\tau^\dagger, \cdot)} \left( \hat{\theta}(\tau^\dagger) \right) | \left\{ \mathbb{U}_{2, 2, \tilde{\tau}_k(\cdot)}^{test, \perp} \left( \hat{\theta}(\tau^\dagger), \tau^\dagger \right) \right\} \right] = \mathbb{U}_{2, 2, \tilde{\tau}_k(\cdot)}^{*, test, \perp} \left( \hat{\theta}(\tau^\dagger), \tau^\dagger \right)$$

where

$$\begin{aligned}
& U_{2,2,\tilde{\tau}_k(\cdot),ij}^{*,test,\perp} \left( \widehat{\theta}(\tau^\dagger), \tau^\dagger \right) \\
&= \left( E_{\widehat{\theta}} \left[ \widehat{\epsilon}_i^2 \widehat{\Delta}_i^2 \right] \right)^{-1} \times \widehat{\epsilon}_i \widehat{\Delta}_i \\
& \left( - \left\{ \begin{aligned} & \left( E_{\widehat{\theta}} \left[ \widehat{\epsilon}_i^2 \widehat{\Delta}_i^2 \right] \right)^{-1} E_{\widehat{\theta}} \left[ \widehat{\epsilon} \widehat{\Delta}^2 \overline{Z}_k^T \right] \\ & \times E_{\widehat{\theta}} \left[ \widehat{\epsilon}^2 \widehat{\Delta} \overline{Z}_k^T \right] \widehat{\epsilon}_j \widehat{\Delta}_j \\ & + E_{\widehat{\theta}} \left[ \widehat{\epsilon}_i^2 \widehat{\Delta}_i \overline{Z}_{k,i}^T \right] \overline{Z}_{k,j} \widehat{\Delta}_j \\ & + E_{\widehat{\theta}} \left[ \widehat{\epsilon}_i \widehat{\Delta}_i^2 \overline{Z}_{k,i}^T \right] \overline{Z}_{k,j} \widehat{\epsilon}_j \end{aligned} \right\} \right)
\end{aligned}$$

As proved in Theorem 41,

$$\left\{ \mathbb{U}_{2,2,\tilde{\tau}_k(\cdot)}^{test,\perp} \left( \theta(\tau^\dagger), \tau^\dagger \right) \right\} = \left\{ \mathbb{V} \left\{ IF_{1,\tau(\cdot),i}^{eff}(\theta) h(O_j; \theta) \right\} : \forall E_\theta [h(O_j; \theta)] = 0 \right\}$$

We assume that

$$\begin{aligned}
& \Pi_{\widehat{\theta}} \left[ \mathbb{IF}_{2,\tilde{\psi}_k(\tau^\dagger,\cdot)} \left( \widehat{\theta}(\tau^\dagger) \right) \mid \left\{ \mathbb{U}_{2,2,\tilde{\tau}_k(\cdot)}^{test,\perp} \left( \widehat{\theta}(\tau^\dagger), \tau^\dagger \right) \right\} \right] \\
&= \mathbb{V} \left\{ IF_{1,\tau(\cdot),i}^{eff}(\theta) h^*(O_j; \theta) \right\},
\end{aligned}$$

then by the definition of the projection, for any  $h(O_j; \theta)$  such that  $E_\theta [h(O_j; \theta)] = 0$ ,

we have

$$\begin{aligned}
& E_{\widehat{\theta}} \left[ \mathbb{IF}_{2,\tilde{\psi}_k(\tau^\dagger,\cdot)} \mathbb{V} \left\{ IF_{1,\tau(\cdot),i}^{eff}(\theta) h(O_j; \theta) \right\} \right] \\
&= E_{\widehat{\theta}} \left[ \mathbb{V} \left\{ IF_{1,\tau(\cdot),i}^{eff}(\theta) h^*(O_j; \theta) \right\} \mathbb{V} \left\{ IF_{1,\tau(\cdot),i}^{eff}(\theta) h(O_j; \theta) \right\} \right],
\end{aligned}$$

which is equivalent to

$$\begin{aligned} & v \left( \hat{\theta} \right)^{-1} \left\{ E_{\hat{\theta}} \left[ \hat{\epsilon}_i^2 \hat{\Delta}_i \bar{Z}_{k,i}^T \bar{Z}_{k,j} \hat{\Delta}_j h(O_j) \right] + E_{\hat{\theta}} \left[ \hat{\epsilon}_i \hat{\Delta}_i^2 \bar{Z}_{k,i}^T \bar{Z}_{k,j} \hat{\epsilon}_j h(O_j) \right] \right\} \\ &= v \left( \hat{\theta} \right)^{-2} \left\{ E_{\hat{\theta}} \left[ \hat{\epsilon}_i^2 \hat{\Delta}_i^2 h^*(O_j) h(O_j) \right] + E_{\hat{\theta}} \left[ \hat{\epsilon}_i \hat{\Delta}_i h(O_i) \hat{\epsilon}_j \hat{\Delta}_j h^*(O_j) \right] \right\}. \end{aligned}$$

As the equation above holds for any mean zero function  $h(O; \theta)$ , therefore

$$\begin{aligned} & \left\{ E_{\hat{\theta}} \left[ \hat{\epsilon}_i^2 \hat{\Delta}_i^2 \right] h^*(O) + \hat{\epsilon} \hat{\Delta} E_{\hat{\theta}} \left[ \hat{\epsilon}_j \hat{\Delta}_j h^*(O_j) \right] \right\} \\ &= v \left( \hat{\theta} \right) \left\{ E_{\hat{\theta}} \left[ \hat{\epsilon}_i^2 \hat{\Delta}_i \bar{Z}_{k,i}^T \right] \bar{Z}_k \hat{\Delta} + E_{\hat{\theta}} \left[ \hat{\epsilon}_i \hat{\Delta}_i^2 \bar{Z}_{k,i}^T \right] \bar{Z}_k \hat{\epsilon} \right\} \end{aligned}$$

$$\Leftrightarrow$$

$$h^*(O)$$

$$= \left( E_{\hat{\theta}} \left[ \hat{\epsilon}_i^2 \hat{\Delta}_i^2 \right] \right)^{-1} \left\{ \begin{array}{c} c_h \left( \hat{\theta} \right) \hat{\epsilon} \hat{\Delta} \\ + v \left( \hat{\theta} \right) E_{\hat{\theta}} \left[ \hat{\epsilon}_i^2 \hat{\Delta}_i \bar{Z}_{k,i}^T \right] \bar{Z}_k \hat{\Delta} \\ + v \left( \hat{\theta} \right) E_{\hat{\theta}} \left[ \hat{\epsilon}_i \hat{\Delta}_i^2 \bar{Z}_{k,i}^T \right] \bar{Z}_k \hat{\epsilon} \end{array} \right\}$$

and  $c_h \left( \hat{\theta} \right)$  is determined by the following equation

$$\begin{aligned} & c_h \left( \hat{\theta} \right) \hat{\epsilon} \hat{\Delta} + v \left( \hat{\theta} \right) E_{\hat{\theta}} \left[ \hat{\epsilon}_i^2 \hat{\Delta}_i \bar{Z}_{k,i}^T \right] \bar{Z}_k \hat{\Delta} \\ &+ v \left( \hat{\theta} \right) E_{\hat{\theta}} \left[ \hat{\epsilon}_i \hat{\Delta}_i^2 \bar{Z}_{k,i}^T \right] \bar{Z}_k \hat{\epsilon} \\ &+ \hat{\epsilon} \hat{\Delta} E_{\hat{\theta}} \left[ \hat{\epsilon} \hat{\Delta} \left( E_{\hat{\theta}} \left[ \hat{\epsilon}_i^2 \hat{\Delta}_i^2 \right] \right)^{-1} \left\{ \begin{array}{c} c_h \left( \hat{\theta} \right) \hat{\epsilon} \hat{\Delta} \\ + v \left( \hat{\theta} \right) E_{\hat{\theta}} \left[ \hat{\epsilon}_i^2 \hat{\Delta}_i \bar{Z}_{k,i}^T \right] \bar{Z}_k \hat{\Delta} \\ + v \left( \hat{\theta} \right) E_{\hat{\theta}} \left[ \hat{\epsilon}_i \hat{\Delta}_i^2 \bar{Z}_{k,i}^T \right] \bar{Z}_k \hat{\epsilon} \end{array} \right\} \right] \\ &= v \left( \hat{\theta} \right) \left\{ E_{\hat{\theta}} \left[ \hat{\epsilon}_i^2 \hat{\Delta}_i \bar{Z}_{k,i}^T \right] \bar{Z}_k \hat{\Delta} + E_{\hat{\theta}} \left[ \hat{\epsilon}_i \hat{\Delta}_i^2 \bar{Z}_{k,i}^T \right] \bar{Z}_k \hat{\epsilon} \right\} \end{aligned}$$



$$\begin{aligned}
& \Leftrightarrow \\
& \widehat{\epsilon} \widehat{\Delta} \left[ \begin{array}{c} c_h \left( \widehat{\theta} \right) + \left( E_{\widehat{\theta}} \left[ \widehat{\epsilon}_i^2 \widehat{\Delta}_i^2 \right] \right)^{-1} \times \\ \left\{ \begin{array}{c} E_{\widehat{\theta}} \left[ \widehat{\epsilon}^2 \widehat{\Delta}^2 \right] c_h \left( \widehat{\theta} \right) \\ + 2v \left( \widehat{\theta} \right) E_{\widehat{\theta}} \left[ \widehat{\epsilon} \widehat{\Delta}^2 \overline{Z}_k^T \right] E_{\widehat{\theta}} \left[ \widehat{\epsilon}^2 \widehat{\Delta} \overline{Z}_k^T \right] \end{array} \right\} \end{array} \right] = 0 \\
& \Leftrightarrow
\end{aligned}$$

$$\begin{aligned}
& c_h \left( \widehat{\theta} \right) \\
& = -v \left( \widehat{\theta} \right) \left( E_{\widehat{\theta}} \left[ \widehat{\epsilon}_i^2 \widehat{\Delta}_i^2 \right] \right)^{-1} E_{\widehat{\theta}} \left[ \widehat{\epsilon} \widehat{\Delta}^2 \overline{Z}_k^T \right] E_{\widehat{\theta}} \left[ \widehat{\epsilon}^2 \widehat{\Delta} \overline{Z}_k^T \right]
\end{aligned}$$

In summary,

$$\begin{aligned}
& \mathbb{U}_{2,2,\widehat{\tau}_k(\cdot)}^{*,test,\perp} \left( \widehat{\theta} \left( \tau^\dagger \right), \tau^\dagger \right) \\
& = \left( E_{\widehat{\theta}} \left[ \widehat{\epsilon}_i^2 \widehat{\Delta}_i^2 \right] \right)^{-1} \times \\
& \quad \mathbb{V} \left\{ \begin{array}{c} \widehat{\epsilon}_i \widehat{\Delta}_i \times \\ \left\{ - \left\{ \begin{array}{c} \left( E_{\widehat{\theta}} \left[ \widehat{\epsilon}_i^2 \widehat{\Delta}_i^2 \right] \right)^{-1} E_{\widehat{\theta}} \left[ \widehat{\epsilon} \widehat{\Delta}^2 \overline{Z}_k^T \right] \\ \times E_{\widehat{\theta}} \left[ \widehat{\epsilon}^2 \widehat{\Delta} \overline{Z}_k^T \right] \widehat{\epsilon}_j \widehat{\Delta}_j \end{array} \right\} \right\} \\ \left\{ \begin{array}{c} + E_{\widehat{\theta}} \left[ \widehat{\epsilon}_i^2 \widehat{\Delta}_i \overline{Z}_{k,i}^T \right] \overline{Z}_{k,j} \widehat{\Delta}_j \\ + E_{\widehat{\theta}} \left[ \widehat{\epsilon}_i \widehat{\Delta}_i^2 \overline{Z}_{k,i}^T \right] \overline{Z}_{k,j} \widehat{\epsilon}_j \end{array} \right\} \end{array} \right\}
\end{aligned}$$

To obtain  $\mathbb{ES}_{2,\widehat{\tau}_k(\cdot)}^{test} \left( \widehat{\theta} \left( \tau^\dagger \right) \right)$ , we divide Eq. (53) by  $var_{\widehat{\theta}} \left\{ \mathbb{ES}_{2,\widehat{\tau}_k(\cdot)}^{test} \left( \widehat{\theta} \left( \tau^\dagger \right) \right) \right\}^{-1}$ .

To obtain  $var_{\widehat{\theta}} \left\{ \mathbb{ES}_{2,\widehat{\tau}_k(\cdot)}^{test} \left( \widehat{\theta} \left( \tau^\dagger \right) \right) \right\}^{-1}$ , we take the variance of both sides of Eq. (53)

under law  $\widehat{\theta}(\tau^\dagger)$  giving

$$\begin{aligned}
& var_{\widehat{\theta}(\tau^\dagger)} \left\{ \mathbb{E} S_{2, \widetilde{\tau}_k(\cdot)}^{test} \left( \widehat{\theta}(\tau^\dagger) \right) \right\}^{-1} \\
&= v \left( \widehat{\theta}(\tau^\dagger) \right)^{-2} var_{\widehat{\theta}(\tau^\dagger)} \left[ \Pi_{\widehat{\theta}(\tau^\dagger)} \left[ \mathbb{IF}_{2, \widetilde{\psi}_k(\tau^\dagger, \cdot)} \left( \widehat{\theta}(\tau^\dagger) \right) | \Gamma_2^{test} \left( \widehat{\theta}(\tau^\dagger), \tau^\dagger \right) \right] \right] \\
&= v \left( \widehat{\theta}(\tau^\dagger) \right)^{-2} \left\{ var_{\widehat{\theta}(\tau^\dagger)} \left[ \mathbb{IF}_{2, \widetilde{\psi}_k(\tau^\dagger, \cdot)} \left( \widehat{\theta}(\tau^\dagger) \right) \right] - var \left[ \mathbb{U}_{2, 2, \widetilde{\tau}_k(\cdot)}^{*, test, \perp} \left( \widehat{\theta}(\tau^\dagger), \tau^\dagger \right) \right] \right\}
\end{aligned}$$

■

**Proof.** (Theorem 44) except part (iii) which was proved in Theorem 23.

Parts (i) and (ii): That  $var_\theta \left[ \mathbb{U}_{2, 2, \widetilde{\tau}_k(\cdot)}^{*, test, \perp} \left( \widehat{\theta}(\tau^\dagger), \tau^\dagger \right) \right] = o_P(1/n)$  and  $var_\theta \left[ \mathbb{Q}_{2, 2, \widetilde{\tau}_k(\cdot)} \left( \widehat{\theta} \right) \right] = o_P(1/n)$  is a straightforward calculation. The remainder of (i) and (ii) follows from the fact that  $var_\theta \left[ \psi_{2, k} \left( \tau, \widehat{\theta} \right) \right] \asymp \max \left( \frac{1}{n}, \frac{k}{n^2} \right)$ .

Part (iv): By part (ii) of Theorem 43, it is sufficient to show that

$$\begin{aligned}
& E_\theta \left[ U_{2, 2, \widetilde{\tau}_k(\cdot)}^{*, test, \perp} \left( \widehat{\theta}(\tau^\dagger), \tau^\dagger \right) \right] \\
&= O_p \left\{ \left( P - \widehat{P} \right) \left( B(\tau^\dagger) - \widehat{B}(\tau^\dagger) \right) \left[ \left( P - \widehat{P} \right) + \left( B(\tau^\dagger) - \widehat{B}(\tau^\dagger) \right) \right] \right\}
\end{aligned}$$

Below we show

$$\begin{aligned}
& E_{\theta} \left[ \mathbb{U}_{2,2,\tilde{\tau}_k(\cdot)}^{*,test,\perp} \left( \hat{\theta}(\tau^\dagger), \tau^\dagger \right) \right] \\
&= \left( E_{\hat{\theta}} \left[ \hat{\epsilon}_i^2 \hat{\Delta}_i^2 \right] \right)^{-1} \times \\
& \left( \begin{array}{c} E_{\theta} \left[ \left( B(\tau^\dagger) - \hat{B}(\tau^\dagger) \right) (P - \hat{P}) \right] \times \\ \left\{ \begin{array}{c} E_{\theta} \left[ (P - \hat{P}) \bar{Z}_k^T \right] E_{\hat{\theta}} \left[ \hat{\epsilon}_i^2 \hat{\Delta}_i \bar{Z}_{k,i}^T \right] \\ + E_{\theta} \left[ \left( B(\tau^\dagger) - \hat{B}(\tau^\dagger) \right) \bar{Z}_k^T \right] E_{\hat{\theta}} \left[ \hat{\epsilon}_i \hat{\Delta}_i^2 \bar{Z}_{k,i}^T \right] \end{array} \right\} \\ - \left\{ \begin{array}{c} \left( E_{\hat{\theta}} \left[ \hat{\epsilon}_i^2 \hat{\Delta}_i^2 \right] \right)^{-1} E_{\hat{\theta}} \left[ \hat{\epsilon} \hat{\Delta}^2 \bar{Z}_k^T \right] \\ \times E_{\hat{\theta}} \left[ \hat{\epsilon}^2 \hat{\Delta} \bar{Z}_k^T \right] \times \\ E_{\theta} \left[ \left( B(\tau^\dagger) - \hat{B}(\tau^\dagger) \right) (P - \hat{P}) \right] \end{array} \right\} \end{array} \right)
\end{aligned}$$

which is

$$O_p \left\{ (P - \hat{P}) (B(\tau^\dagger) - \hat{B}(\tau^\dagger)) \left[ (P - \hat{P}) + (B(\tau^\dagger) - \hat{B}(\tau^\dagger)) \right] \right\}$$

when, as is the case under our assumptions  $E_{\hat{\theta}(\tau^\dagger)} \left[ \left\{ Y(\tau^\dagger) - \hat{B}(\tau^\dagger) \right\} \left\{ A - \hat{P} \right\}^2 \bar{Z}_k^T \right]$  and  $E_{\hat{\theta}(\tau^\dagger)} \left[ \left\{ Y(\tau^\dagger) - \hat{B}(\tau^\dagger) \right\} \left\{ A - \hat{P} \right\}^2 \bar{Z}_k^T \right]$  are both order  $O_p(1)$ , but would be

$$O_p \left\{ (P - \hat{P}) (B(\tau^\dagger) - \hat{B}(\tau^\dagger)) \left[ (P - \hat{P})^2 + (B(\tau^\dagger) - \hat{B}(\tau^\dagger))^2 \right] \right\}$$

in the (unlikely) special case in which the semiparametric regression model was precisely true, since then

$$E_{\hat{\theta}} [Y(\tau^\dagger) A | X] / \{ E_{\hat{\theta}} [Y(\tau^\dagger) | X] E_{\hat{\theta}} [A | X] \} = 1 + o_p(1)$$

so

$$E_{\hat{\theta}(\tau^\dagger)} \left[ \left\{ Y(\tau^\dagger) - \hat{B}(\tau^\dagger) \right\}^2 \left\{ A - \hat{P} \right\} \bar{Z}_k^T \right] = O_p(P - \hat{P})$$

and

$$E_{\widehat{\theta}(\tau^\dagger)} \left[ \left\{ Y(\tau^\dagger) - \widehat{B}(\tau^\dagger) \right\} \left\{ A - \widehat{P} \right\}^2 \overline{Z}_k^T \right] = O_p \left( B(\tau^\dagger) - \widehat{B} \right).$$

The expression for  $E_\theta \left[ \mathbb{U}_{2,2,\widetilde{\tau}_k(\cdot)}^{*,test,\perp} \left( \widehat{\theta}(\tau^\dagger), \tau^\dagger \right) \right]$  is obtained from the formula for  $U_{2,2,\widetilde{\tau}_k(\cdot)}^{*,test,\perp} \left( \widehat{\theta}(\tau^\dagger), \tau^\dagger \right)$  in Theorem 43 by noting that, because  $\mathbb{V} \left[ \left( Y(\tau^\dagger) - \widehat{B}(\tau^\dagger) \right) \left( A - \widehat{P} \right) \right] = \psi_{1,k} \left( \tau^\dagger, \widehat{\theta}(\tau^\dagger) \right)$

and

$$\begin{aligned} \widetilde{\psi}_k(\tau^\dagger, \theta) &= 0, \\ E_\theta \left[ \left( Y(\tau^\dagger) - \widehat{B}(\tau^\dagger) \right) \left( A - \widehat{P} \right) \right] \\ &= E_\theta \left[ \psi_{1,k} \left( \tau^\dagger, \widehat{\theta}(\tau^\dagger) \right) - \widetilde{\psi}_k(\tau^\dagger, \theta) \right] \\ &= E_\theta \left[ \left( P - \widehat{P} \right) \overline{Z}_k^T \right] E_\theta \left[ \overline{Z}_k \overline{Z}_k^T \right] E_\theta \left[ \left( B(\tau^\dagger) - \widehat{B}(\tau^\dagger) \right) \overline{Z}_k \right] \end{aligned}$$

by Theorem 23.

Part (v): We first note that by Theorem 2,  $\frac{E_\theta[\tau_{2,k}(\hat{\theta}) - \tilde{\tau}_k(\theta)]}{E_\theta[\mathbb{IF}_{3,3,\tilde{\tau}_k(\cdot)}(\hat{\theta})]} = -(1 + o_p(1))$ . It

can be shown that

$$\mathbb{IF}_{3,3,\tilde{\tau}_k(\cdot)}(\theta) = (-\psi_\tau)^{-1} \times \mathbb{V} \left\{ \begin{aligned} & IF_{3,3,\psi(\tau^\dagger, \cdot), i_1 i_2 i_3}(\theta) + \frac{1}{6} \psi_{\setminus \tau^3} IF_{1,\tau(\cdot), i_1}(\theta) IF_{1,\tau(\cdot), i_2}(\theta) IF_{1,\tau(\cdot), i_3}(\theta) \\ & + \frac{1}{3} \psi_{\setminus \tau^2} \begin{pmatrix} IF_{1,\tau(\cdot), i_1}(\theta) IF_{2,2,\tau(\cdot), i_2 i_3}(\theta) \\ + IF_{1,\tau(\cdot), i_2}(\theta) IF_{2,2,\tau(\cdot), i_1 i_3}(\theta) \\ + IF_{1,\tau(\cdot), i_3}(\theta) IF_{2,2,\tau(\cdot), i_1 i_2}(\theta) \end{pmatrix} \\ & + \frac{1}{3} \begin{pmatrix} d_{1,\theta} \left( \frac{\partial IF_{1,\psi(\tau^\dagger, \cdot), i_1}(\theta)}{\partial \tau} \right) IF_{2,2,\tau(\cdot), i_2 i_3}(\theta) \\ + d_{1,\theta} \left( \frac{\partial IF_{1,\psi(\tau^\dagger, \cdot), i_2}(\theta)}{\partial \tau} \right) IF_{2,2,\tau(\cdot), i_1 i_3}(\theta) \\ + d_{1,\theta} \left( \frac{\partial IF_{1,\psi(\tau^\dagger, \cdot), i_3}(\theta)}{\partial \tau} \right) IF_{2,2,\tau(\cdot), i_1 i_2}(\theta) \end{pmatrix} \\ & + \frac{1}{3} \begin{pmatrix} IF_{1,\tau(\cdot), i_1}(\theta) d_{2,\theta} \left( \frac{\partial IF_{2,2,\psi(\tau^\dagger, \cdot), i_2 i_3}(\theta)}{\partial \tau} \right) \\ + IF_{1,\tau(\cdot), i_2}(\theta) d_{2,\theta} \left( \frac{\partial IF_{2,2,\psi(\tau^\dagger, \cdot), i_1 i_3}(\theta)}{\partial \tau} \right) \\ + IF_{1,\tau(\cdot), i_3}(\theta) d_{2,\theta} \left( \frac{\partial IF_{2,2,\psi(\tau^\dagger, \cdot), i_1 i_2}(\theta)}{\partial \tau} \right) \end{pmatrix} \\ & + \frac{1}{6} \begin{pmatrix} d_{1,\theta} \left( \frac{\partial^2 IF_{1,\psi(\tau^\dagger, \cdot), i_1}(\theta)}{\partial \tau^2} \right) IF_{1,\tau(\cdot), i_2}(\theta) IF_{1,\tau(\cdot), i_3}(\theta) \\ + d_{1,\theta} \left( \frac{\partial^2 IF_{1,\psi(\tau^\dagger, \cdot), i_2}(\theta)}{\partial \tau^2} \right) IF_{1,\tau(\cdot), i_1}(\theta) IF_{1,\tau(\cdot), i_3}(\theta) \\ + d_{1,\theta} \left( \frac{\partial^2 IF_{1,\psi(\tau^\dagger, \cdot), i_3}(\theta)}{\partial \tau^2} \right) IF_{1,\tau(\cdot), i_1}(\theta) IF_{1,\tau(\cdot), i_2}(\theta) \end{pmatrix} \end{aligned} \right\}$$

From the fact that  $\partial^3 \tilde{\psi}_k(\tau, \theta) / \partial \tau^3 = \partial^2 \tilde{\psi}_k(\tau, \theta) / \partial \tau^2 = \partial^2 IF_{1, \tilde{\psi}_k(\tau, \cdot)}(\hat{\theta}) / \partial \tau^2 = 0$ ,

we conclude that the order of  $E_\theta \left[ \tau_{2,k}(\hat{\theta}) - \tilde{\tau}_k(\theta) \right]$  is equal to the order of

$$\begin{aligned} & E_\theta \left[ IF_{3,3, \tilde{\psi}_k(\tau, \cdot)}(\hat{\theta}) \right] + E_\theta \left[ d_{1,\theta} \left( \partial IF_{1, \tilde{\psi}_k(\tau, \cdot)}(\hat{\theta}) / \partial \tau \right) \right] E_\theta \left[ IF_{2,2, \tilde{\tau}_k(\cdot)}(\hat{\theta}) \right] \\ & + E_\theta \left[ d_{2,\theta} \left( \partial IF_{2,2, \tilde{\psi}_k(\tau, \cdot)}(\hat{\theta}) / \partial \tau \right) \right] E_\theta \left[ IF_{1, \tilde{\tau}_k(\cdot)}(\hat{\theta}) \right] \end{aligned}$$

Now

$$\begin{aligned} E_\theta \left[ IF_{3,3, \tilde{\psi}_k(\tau^\dagger, \cdot)}(\hat{\theta}) \right] &= O_p \left[ \left( P - \hat{P} \right) \left( B - \hat{B} \right) \left( \frac{g(X)}{\hat{g}(X)} - 1 \right) \right], \\ E_\theta \left[ d_{1,\theta} \left( \partial IF_{1, \tilde{\psi}_k(\tau, \cdot)}(\hat{\theta}) / \partial \tau \right) \right] &= E_\theta \left[ \left( A - \hat{P} \right)^2 \right] - E_{\hat{\theta}} \left[ \left( A - \hat{P} \right)^2 \right] \\ &= E_{\hat{\theta}} \left[ \left( \frac{f(A|X)}{\hat{f}(A|X)} \frac{g(X)}{\hat{g}(X)} - 1 \right) \left( A - \hat{P} \right)^2 \right] \\ &= E_{\hat{\theta}} \left[ \left( \left[ \left( \frac{f(A|X)}{\hat{f}(A|X)} - 1 \right) \frac{g(X)}{\hat{g}(X)} \right] \right) \left( A - \hat{P} \right)^2 \right] \\ &+ E_{\hat{\theta}} \left[ \left( \frac{g(X)}{\hat{g}(X)} - 1 \right) \left( A - \hat{P} \right)^2 \right] = O_p \left[ \left( P - \hat{P} \right) + \left( \frac{g(X)}{\hat{g}(X)} - 1 \right) \right] \end{aligned}$$

by  $A$  binary,

$$\begin{aligned} E_\theta \left[ \partial IF_{2,2, \tilde{\psi}_k(\tau, \cdot)}(\hat{\theta}) / \partial \tau \right] &= - \left\{ E_\theta \left[ \left( A - \hat{P} \right) \right] \right\}^2 = O_p \left[ \left( P - \hat{P} \right)^2 \right] \\ E_\theta \left[ IF_{1, \tilde{\tau}_k(\cdot)}(\hat{\theta}) \right] &= E_\theta \left[ \left( A - \hat{P} \right) \left( Y - \hat{B} \right) \right] \\ &= E_{\hat{\theta}} \left[ \left( P - \hat{P} \right) \left( B - \hat{B} \right) \right] \end{aligned}$$

and using  $\partial^2 \tilde{\psi}_k(\tau, \theta) / \partial \tau^2 = 0$  and the explicit expression for  $\mathbb{IF}_{2,2,\tilde{\tau}_k(\cdot)}(\theta)$  in Eq (52)

$$\begin{aligned}
& E_\theta \left[ IF_{2,2,\tilde{\tau}_k(\cdot)}(\hat{\theta}) \right] \\
&= O_p \left[ E_\theta \left[ IF_{2,2,\tilde{\psi}_k(\tau^\dagger, \cdot)}(\hat{\theta}) \right] \right] + E_\theta \left[ \partial IF_{1,\tilde{\psi}_k(\tau, \cdot)}(\hat{\theta}) / \partial \tau \right] \\
&\times E_\theta \left[ IF_{1,\tilde{\tau}_k(\cdot)}(\hat{\theta}) \right] \\
&= O_p \left[ (P - \hat{P}) (B - \hat{B}) \right] \\
&+ O_p \left[ (P - \hat{P}) + \left( \frac{g(X)}{\hat{g}(X)} - 1 \right) \right] \times \\
&O_p \left[ (P - \hat{P}) + \left( \frac{g(X)}{\hat{g}(X)} - 1 \right) + (B - \hat{B}) \right]
\end{aligned}$$

Combining terms completes the proof. ■

Next, we motivate and derive the formula of  $\mathbb{IF}_{m,m,\tau(\cdot)}(\theta)$  for an assumed unique functional  $\tau(\theta)$  implicitly defined by  $0 = \psi(\tau(\theta), \theta)$ ,  $\theta \in \Theta$ . To motivate the general formula of  $\mathbb{IF}_{m,m,\tau(\cdot)}(\theta)$  for arbitrary  $m$ , we first consider the following formula for  $\mathbb{IF}_{44,\tau(\cdot)}$  which was derived from  $\mathbb{IF}_{1,\tau(\cdot)}$  following part 5c) of Theorem 3.

$$\begin{aligned}
& -\psi_{\setminus\tau} \mathbb{IF}_{44,\tau(\cdot)} \\
& = \mathbb{IF}_{4,4,\psi(\tau,\cdot)} \\
& + \frac{1}{24} \mathbb{V} \left\{ \begin{aligned} & \psi_{\setminus\tau^4} IF_{1,\tau}(i) IF_{1,\tau}(j) IF_{1,\tau}(s) IF_{1,\tau}(t) \\ & + d_{1,\theta} \left( \frac{\partial IF_{1,\tau}(i)}{\partial \tau^3} \right) IF_{1,\tau}(j) IF_{1,\tau}(s) IF_{1,\psi(\tau,\cdot)}(t) \\ & + IF_{1,\tau}(i) d_{1,\theta} \left( \frac{\partial IF_{1,\tau}(j)}{\partial \tau^3} \right) IF_{1,\tau}(s) IF_{1,\psi(\tau,\cdot)}(t) \\ & + IF_{1,\tau}(i) IF_{1,\tau}(j) d_{1,\theta} \left( \frac{\partial IF_{1,\tau}(s)}{\partial \tau^3} \right) IF_{1,\psi(\tau,\cdot)}(t) \\ & + IF_{1,\tau}(i) IF_{1,\tau}(j) IF_{1,\tau}(s) d_{1,\theta} \left( \frac{\partial IF_{1,\psi(\tau,\cdot)}(t)}{\partial \tau^3} \right) \end{aligned} \right\} \\
& + \frac{1}{2} \mathbb{V} \left\{ \begin{aligned} & \psi_{\setminus\tau^3} IF_{22,\tau}(ij) IF_{1,\tau}(s) IF_{1,\tau}(t) \\ & + d_{1,\theta} \left( \frac{\partial IF_{1,\psi}(t)}{\partial \tau^2} \right) IF_{22,\tau}(ij) IF_{1,\tau}(s) \\ & + d_{1,\theta} \left( \frac{\partial IF_{1,\psi}(s)}{\partial \tau^2} \right) IF_{22,\tau}(ij) IF_{1,\tau}(t) \\ & + d_{1,\theta} \left( \frac{\partial IF_{22,\psi}(ij)}{\partial \tau^2} \right) IF_{1,\tau}(t) IF_{1,\tau}(s) \end{aligned} \right\} \\
& + \frac{1}{2} \mathbb{V} \left\{ \begin{aligned} & \psi_{\setminus\tau^2} IF_{22,\tau}(ij) IF_{22,\tau}(st) + \\ & d_{2,\theta} \left( \frac{\partial IF_{22,\psi}(ij)}{\partial \tau} \right) IF_{22,\tau}(st) \\ & + IF_{22,\tau}(ij) d_{2,\theta} \left( \frac{\partial IF_{22,\psi}(st)}{\partial \tau} \right) \end{aligned} \right\} \\
& + \mathbb{V} \left\{ \begin{aligned} & \psi_{\setminus\tau^2} IF_{1,\tau}(i) IF_{33,\tau}(jst) + \\ & d_{1,\theta} \left( \frac{\partial IF_{1,\psi}(i)}{\partial \tau} \right) IF_{33,\tau}(jst) + d_{3,\theta} \left( \frac{\partial IF_{33,\psi}(jst)}{\partial \tau} \right) IF_{1,\tau}(i) \end{aligned} \right\}
\end{aligned}$$

This formula for  $-\psi_{\setminus\tau} \mathbb{IF}_{44,\tau(\cdot)}$  reveals a very nice pattern. Note that in addition to

$\mathbb{IF}_{4,4,\psi(\tau,\cdot)}$ , the RHS consists of four pieces with leading terms  $\psi_{\setminus\tau^4} IF_{1,\tau}(i) IF_{1,\tau}(j) IF_{1,\tau}(s) IF_{1,\tau}(t)$ ,



$\psi_{\setminus\tau^3} IF_{22,\tau}(ij) IF_{1,\tau}(s) IF_{1,\tau}(t)$ ,  $\psi_{\setminus\tau^2} IF_{22,\tau}(ij) IF_{22,\tau}(st)$ , and  $\psi_{\setminus\tau^2} IF_{1,\tau}(i) IF_{33,\tau}(jst)$  respectively. Within each piece, the remaining terms can be constructed simply by applying the algorithm **TE** described below.

**Algorithm 59 TE** *i) Remove the partial derivative of  $\psi$  wrt.  $\tau$ ; ii) for each factor of the leading term, replace the  $\tau(\cdot)$  in the subscript with  $\psi(\tau, \cdot)$ ; and iii) partially differentiate the newly replaced factor wrt.  $\tau$ , and make the partial derivative degenerate. Here the order of the partial derivative equals the total number of the factors in the leading term minus 1.*

Moreover, each piece corresponds to one of the following ways of representing the number 4 as a sum:  $1 + 1 + 1 + 1$ ,  $1 + 1 + 2$ ,  $2 + 2$ , and  $1 + 3$ . Furthermore, assume  $m$  can be written as  $m = \sum_{r=1}^{m-1} \varkappa_{m,r} \times r$  with  $\varkappa_{m,r} \geq 0$ , e.g.,  $4 = 4 \times 1 + 0 \times 2 + 0 \times 3$ , then the number in front of the piece corresponding to the sum representation  $(\varkappa_{m,1}, \varkappa_{m,2}, \dots, \varkappa_{m,m-1})$  equals  $\left( \prod_{r=1}^{m-1} \varkappa_{m,r}! \right)^{-1}$ . Note that  $0! = 1$ .

Now we are ready to generalize this expression to arbitrary  $m$ , and prove it by induction.

**Lemma 60** *Let  $\tau(\theta)$  be the assumed unique functional defined by  $0 = \psi(\tau(\theta), \theta)$ ,  $\theta \in$*

$\Theta$ . Then, for  $\theta \in \Theta(\tau^\dagger)$ , whenever  $\mathbb{IF}_{m,\psi(\tau^\dagger,\cdot)}(\theta)$  and  $\mathbb{IF}_{m,\tau(\cdot)}(\theta)$  exist ,

$$\mathbb{IF}_{1,\tau(\cdot)}(\theta) = -\psi_{\setminus\tau}^{-1} \mathbb{IF}_{1,\psi(\tau,\cdot)}(\theta) \quad (73)$$

$$\begin{aligned} & -\psi_{\setminus\tau} \mathbb{IF}_{m,m,\tau(\cdot)}(\theta) \\ & = \mathbb{IF}_{m,m,\psi(\tau,\cdot)}(\theta) \\ & + \sum_{(\varkappa_{m,1}, \varkappa_{m,2}, \dots, \varkappa_{m,m-1})} \left( \prod_{r=1}^{m-1} \varkappa_{m,r}! \right)^{-1} \mathbb{V} \left\{ \begin{aligned} & \psi_{\setminus\tau^{sum(\bar{\varkappa}_m)}} \prod_{r=1}^{m-1} \left( \prod_{s=1}^{\varkappa_{m,r}} IF_{r,r,\tau(\cdot), \bar{i}_{r,s}}^{(s)}(\theta) \right) \\ & + \sum_{r,s}^{m-1, \varkappa_{m,r}} \left( d_{r,\theta} \left( \frac{\partial^{[sum(\bar{\varkappa}_m)-1]} \left( IF_{r,r,\psi(\tau,\cdot), \bar{i}_{r,s}}^{(s)} \right)}{\partial \tau^{[sum(\bar{\varkappa}_m)-1]}} \right) \times \right. \\ & \left. \prod_{(r_1, s_1) \neq (r,s)} IF_{r_1, r_1, \tau(\cdot), \bar{i}_{r_1, s_1}}^{(s)}(\theta) \right) \end{aligned} \right\} \end{aligned} \quad (74)$$

where

$$\bar{\varkappa}_m \equiv (\varkappa_{m,1}, \varkappa_{m,2}, \dots, \varkappa_{m,m-1}), \text{ and } m = \sum_{r=1}^{m-1} \varkappa_{m,r} \times r, \quad \varkappa_{m,r} \geq 0 \quad \forall 1 \leq r \leq m-1$$

$$sum(\bar{\varkappa}_m) \equiv \sum_{r=1}^{m-1} \varkappa_{m,r}$$

$$\bar{i}_{r,s} \equiv (i_{l_{r,s}+1}, \dots, i_{l_{r,s}+r}) \quad \text{where } l_{r,s} \equiv \sum_{q=1}^{r-1} \varkappa_{m,q} \times q + (s-1)r$$

Note that

$$\mathbb{V} \left\{ \sum_{r,s}^{m-1, \varkappa_{m,r}} \left( d_{r,\theta} \left( \frac{\partial^{[sum(\bar{\varkappa}_m)-1]} \left( IF_{r,r,\psi(\tau,\cdot), \bar{i}_{r,s}}^{(s)} \right)}{\partial \tau^{[sum(\bar{\varkappa}_m)-1]}} \right) \times \prod_{(r_1, s_1) \neq (r,s)} IF_{r_1, r_1, \tau(\cdot), \bar{i}_{r_1, s_1}}^{(s)}(\theta) \right) \right\}$$

can be constructed by applying Algorithm **TE** to the leading term

$$\mathbb{V} \left\{ \psi_{\setminus\tau^{sum(\bar{\varkappa}_m)}} \prod_{r=1}^{m-1} \left( \prod_{s=1}^{\varkappa_{m,r}} IF_{r,r,\tau(\cdot), \bar{i}_{r,s}}^{(s)}(\theta) \right) \right\}.$$

**Proof.** Eq. (73) has been proved in Theorem 42. Next, we shall prove eq. (74) by induction. The case where  $m = 2$  was proved in Theorem 42 as well. Now, we assume eq. (74) holds for  $m$  and prove it is also true for  $m+1$  by part 5c) of Theorem 3. Specifically, by induction assumption,

$$\begin{aligned}
& -\psi_{\setminus\tau} if_{m,m,\tau(\cdot)}^{(s)}(\mathbf{o}_{i_m}; \theta) \\
& = if_{m,m,\psi(\tau,\cdot)}^{(s)}(\mathbf{o}_{i_m}; \theta) + \\
& \quad \sum_{(\varkappa_{m,1}, \varkappa_{m,2}, \dots, \varkappa_{m,m-1})} \left( \prod_{r=1}^{m-1} \varkappa_{m,r}! \right)^{-1} \times \\
& \quad \left\{ \begin{aligned} & \psi_{\setminus\tau^{sum(\overline{\varkappa}_m)}} \prod_{r=1}^{m-1} \left( \prod_{s=1}^{\varkappa_{m,r}} IF_{r,r,\tau(\cdot), \bar{i}_{r,s}}^{(s)}(\theta) \right) \\ & + \sum_{r,s}^{m-1, \varkappa_{m,r}} \left( d_{r,\theta} \left( \frac{\partial^{[sum(\overline{\varkappa}_m)-1]} \left( IF_{r,r,\psi(\tau,\cdot), \bar{i}_{r,s}}^{(s)} \right)}{\partial \tau^{[sum(\overline{\varkappa}_m)-1]}} \right) \times \right. \\ & \quad \left. \prod_{(r_1, s_1) \neq (r,s)} IF_{r_1, r_1, \tau(\cdot), \bar{i}_{r_1, s_1}}^{(s)}(\theta) \right) \end{aligned} \right\}
\end{aligned}$$

Consider any sufficiently smooth 1-dimensional parametric submodel  $\theta_t$  mapping  $R$  to  $\Theta$ . For any  $\theta$  in the range of  $\theta_t$ , differentiate both sides of the above equation

w.r.t  $t$  and evaluate at  $t^* \equiv \theta_t^{-1}(\theta)$ , then

$$\begin{aligned}
& -\psi_{\setminus\tau} if_{m,m,\tau(\cdot),\setminus t}^{(s)}(\mathbf{o}_{i_m};\theta) - \left( \psi_{\setminus\tau^2\tau\setminus t}(\theta) + \frac{\partial}{\partial t} \psi_{\setminus\tau}(\tau, \theta_t) \Big|_{t=t^*} \right) if_{m,m,\tau(\cdot)}^{(s)}(\mathbf{o}_{i_m};\theta) \\
& = \frac{\partial IF_{m,m,\psi(\tau,\cdot)}^{(s)}(\tau, \theta)}{\partial \tau} \tau_{\setminus t}(\theta) + \frac{\partial IF_{m,m,\psi(\tau,\cdot)}^{(s)}(\tau, \theta_t)}{\partial t} \Big|_{t=t^*} \\
& + \sum_{(\varkappa_{m,1}, \varkappa_{m,2}, \dots, \varkappa_{m,m-1})} \left( \prod_{r=1}^{m-1} \varkappa_{m,r}! \right)^{-1} \times \\
& \left\{ \begin{aligned} & \left[ \begin{aligned} & \psi_{\setminus\tau^{sum(\overline{\varkappa}_m)+1}\tau\setminus t}(\theta) \\ & + \frac{\partial \psi_{\setminus\tau^{sum(\overline{\varkappa}_m)}(\tau, \theta_t)}{\partial t} \Big|_{t=t^*} \end{aligned} \right] \prod_{r=1}^{m-1} \left( \prod_{s=1}^{\varkappa_{m,r}} IF_{r,r,\tau(\cdot),\bar{i}_{r,s}}^{(s)}(\theta) \right) \\ & + \psi_{\setminus\tau^{sum(\overline{\varkappa}_m)}} \sum_{(r,s)}^{sum(\overline{\varkappa}_m)} IF_{r,r,\tau(\cdot),\bar{i}_{r,s},\setminus t}^{(s)}(\theta) \prod_{(r_1,s_1) \neq (r,s)} IF_{r_1,r_1,\tau(\cdot),\bar{i}_{r_1,s_1}}^{(s)}(\theta) \\ & \times \left[ \begin{aligned} & \frac{\partial^{sum(\overline{\varkappa}_m)} \left( IF_{r,r,\psi(\tau,\cdot),\bar{i}_{r,s}}^{(s)} \right)}{\partial \tau^{sum(\overline{\varkappa}_m)}} \tau_{\setminus t}(\theta) \\ & + \frac{\partial^{sum(\overline{\varkappa}_m)} \left( IF_{r,r,\psi(\tau,\cdot),\bar{i}_{r,s}}^{(s)}(\tau, \theta_t) \right)}{\partial \tau^{[sum(\overline{\varkappa}_m)-1]}} \Big|_{t=t^*} \end{aligned} \right] \\ & \times \prod_{(r_1,s_1) \neq (r,s)}^{sum(\overline{\varkappa}_m)} IF_{r_1,r_1,\tau(\cdot),\bar{i}_{r_1,s_1}}^{(s)}(\theta) \\ & + \frac{\partial^{[sum(\overline{\varkappa}_m)-1]} \left( IF_{r,r,\psi(\tau,\cdot),\bar{i}_{r,s}}^{(s)} \right)}{\partial \tau^{[sum(\overline{\varkappa}_m)-1]}} \\ & \times \sum_{(r_1,s_1) \neq (r,s)} IF_{r_1,r_1,\tau(\cdot),\bar{i}_{r_1,s_1},\setminus t}^{(s)}(\theta) \\ & \times \prod_{(r_2,s_2) \neq (r_1,s_1) \neq (r,s)} IF_{r_2,r_2,\tau(\cdot),\bar{i}_{r_2,s_2}}^{(s)}(\theta) \end{aligned} \right\} \quad (75)
\end{aligned}$$

Note that  $\frac{\partial^p \left( IF_{r,r,\psi(\tau,\cdot),\bar{i}_{r,s}}^{(s)}(\tau, \theta_t) \right)}{\partial \tau^{p-1} \partial t} \Big|_{t=t^*}$  is the derivative of  $\frac{\partial^{p-1} \left( IF_{r,r,\psi(\tau,\cdot),\bar{i}_{r,s}}^{(s)} \right)}{\partial \tau^{p-1}}$  w.r.t.  $t$

while fixing  $\tau$  at  $\tau(\theta)$ . Therefore,

$$\begin{aligned}
& \frac{\partial^p \left( IF_{r,r,\psi(\tau,\cdot),\bar{i}_{r,s}}^{(s)}(\tau, \theta_t) \right)}{\partial \tau^{p-1} \partial t} \Big|_{t=t^*} \\
&= \frac{\partial^{p-1}}{\partial \tau^{p-1}} \left( \frac{\partial}{\partial t} IF_{r,r,\psi(\tau,\cdot),\bar{i}_{r,s}}^{(s)}(\tau, \theta_t) \Big|_{t=t^*} \right) \\
&= \frac{\partial^{p-1}}{\partial \tau^{p-1}} E_\theta \left[ if_{1, IF_{r,r,\psi(\tau,\cdot),\bar{i}_{r,s}}^{(s)}}(\cdot)(O_{m+1}; \theta) s_{1,t}(O_{m+1}) \right] \\
&= E_\theta \left[ \frac{\partial^{p-1} if_{1, IF_{r,r,\psi(\tau,\cdot),\bar{i}_{r,s}}^{(s)}}(\cdot)(O_{m+1}; \theta)}{\partial \tau^{p-1}} s_{1,t}(O_{m+1}) \right],
\end{aligned}$$

and

$$\begin{aligned}
& - \psi_{\setminus \tau} i f_{1, i f_{m, m, \tau(\cdot), \setminus t}^{(s)}(\mathbf{o}_{i_m}; \cdot)}^{(s)}(O_{i_{m+1}}; \theta) \\
& = i f_{1, i f_{m, m, \psi(\tau, \cdot)}^{(s)}(\mathbf{o}_{i_m}; \cdot)}^{(s)}(O_{i_{m+1}}; \theta) \\
& + \left\{ \begin{aligned} & \psi_{\setminus \tau^2} I F_{1, \tau(\cdot), i_{m+1}}(\theta) i f_{m, m, \tau(\cdot)}^{(s)}(\mathbf{o}_{i_m}; \theta) \\ & + \frac{\partial}{\partial \tau} I F_{1, \psi(\tau, \cdot), i_{m+1}} i f_{m, m, \tau(\cdot)}^{(s)}(\mathbf{o}_{i_m}; \theta) \\ & + I F_{m, m, \psi(\tau, \cdot), \setminus \tau}^{(s)} I F_{1, \tau(\cdot), i_{m+1}} \end{aligned} \right\} \\
& + \sum_{(\varkappa_{m,1}, \varkappa_{m,2}, \dots, \varkappa_{m,m-1})} \left( \prod_{r=1}^{m-1} \varkappa_{m,r}! \right)^{-1} \times \\
& \left\{ \begin{aligned} & \left\{ \begin{aligned} & \left[ \begin{aligned} & \psi_{\setminus \tau^{sum(\overline{\varkappa}_m)+1}} I F_{1, \tau(\cdot), i_{m+1}}(\theta) \\ & + \frac{\partial^{sum(\overline{\varkappa}_m)}}{\partial \tau^{sum(\overline{\varkappa}_m)}} (I F_{1, \psi(\tau, \cdot), i_{m+1}}) \\ & \times \prod_{r=1}^{m-1} \left( \prod_{s=1}^{\varkappa_{m,r}} I F_{r, r, \tau(\cdot), \bar{i}_{r,s}}^{(s)}(\theta) \right) \end{aligned} \right] \\ & + \sum_{(r,s)}^{sum(\overline{\varkappa}_m)} \left[ \begin{aligned} & \frac{\partial^{sum(\overline{\varkappa}_m)} (I F_{r, r, \psi(\tau, \cdot), \bar{i}_{r,s}}^{(s)})}{\partial \tau^{sum(\overline{\varkappa}_m)}} I F_{1, \tau(\cdot), i_{m+1}}(\theta) \\ & \times \prod_{(r_1, s_1) \neq (r, s)} I F_{r_1, r_1, \tau(\cdot), \bar{i}_{r_1, s_1}}^{(s)}(\theta) \end{aligned} \right] \end{aligned} \right\} \\
& + \sum_{(r,s)}^{sum(\overline{\varkappa}_m)} \left\{ \begin{aligned} & \left[ \begin{aligned} & \psi_{\setminus \tau^{sum(\overline{\varkappa}_m)}} i f_{1, I F_{r, r, \tau(\cdot), \bar{i}_{r,s}}^{(s)}(\cdot)}^{(s)}(O_{i_{m+1}}; \theta) \\ & \times \prod_{(r_1, s_1) \neq (r, s)} I F_{r_1, r_1, \tau(\cdot), \bar{i}_{r_1, s_1}}^{(s)}(\theta) \end{aligned} \right] \\ & + \left[ \begin{aligned} & \frac{\partial^{sum(\overline{\varkappa}_m)} (i f_{1, I F_{r, r, \psi(\tau, \cdot), \bar{i}_{r,s}}^{(s)}(\cdot)}^{(s)}(O_{i_{m+1}}; \theta))}{\partial \tau^{[sum(\overline{\varkappa}_m)-1]}} \\ & \times \prod_{(r_1, s_1) \neq (r, s)} I F_{r_1, r_1, \tau(\cdot), \bar{i}_{r_1, s_1}}^{(s)}(\theta) \end{aligned} \right] \end{aligned} \right\} \\
& + \sum_{(r,s)}^{sum(\overline{\varkappa}_m)} \left\{ \begin{aligned} & \sum_{(r_1, s_1) \neq (r, s)} \left( \begin{aligned} & \frac{\partial^{[sum(\overline{\varkappa}_m)-1]} (I F_{r, r, \psi(\tau, \cdot), \bar{i}_{r,s}}^{(s)})}{\partial \tau^{[sum(\overline{\varkappa}_m)-1]}} \times \\ & i f_{1, I F_{r_1, r_1, \tau(\cdot), \bar{i}_{r_1, s_1}}^{(s)}(\cdot)}^{(s)}(O_{i_{m+1}}; \theta) \times \\ & \prod_{(r_2, s_2) \neq (r_1, s_1) \neq (r, s)} I F_{r_2, r_2, \tau(\cdot), \bar{i}_{r_2, s_2}}^{(s)}(\theta) \end{aligned} \right) \end{aligned} \right\} \end{aligned} \right\}
\end{aligned}$$

Consider the last term

$$\sum_{(r,s)}^{sum(\overline{\alpha}_m)} \left\{ \frac{\partial^{[sum(\overline{\alpha}_m)-1]} \left( IF_{r,r,\psi(\tau,\cdot),\bar{i}_{r,s}}^{(s)} \right)}{\partial \tau^{[sum(\overline{\alpha}_m)-1]}} \sum_{(r_1,s_1) \neq (r,s)} if_{1,IF_{r_1,r_1,\tau(\cdot),\bar{i}_{r_1,s_1}}^{(s)}}^{(s)}(\cdot) (O_{i_{m+1}}; \theta) \right. \\ \left. \times \prod_{(r_2,s_2) \neq (r_1,s_1) \neq (r,s)} IF_{r_2,r_2,\tau(\cdot),\bar{i}_{r_2,s_2}}^{(s)}(\theta) \right\}$$

WLOG, we exchange  $(r, s)$  with  $(r_1, s_1)$ , then we have

$$\sum_{(r,s) \neq (r_1,s_1)}^{sum(\overline{\alpha}_m), sum(\overline{\alpha}_m)} \left[ \frac{\partial^{[sum(\overline{\alpha}_m)-1]} \left( IF_{r_1,r_1,\psi(\tau,\cdot),\bar{i}_{r_1,s_1}}^{(s)} \right)}{\partial \tau^{[sum(\overline{\alpha}_m)-1]}} \times \right. \\ \left. if_{1,IF_{r,r,\tau(\cdot),\bar{i}_{r,s}}^{(s)}}^{(s)}(\cdot) (O_{i_{m+1}}; \theta) \prod_{(r_2,s_2) \neq (r_1,s_1) \neq (r,s)} IF_{r_2,r_2,\tau(\cdot),\bar{i}_{r_2,s_2}}^{(s)}(\theta) \right]$$

and the sum of the last two terms equals

$$\sum_{(r,s)}^{sum(\overline{\alpha}_m)} \left( \begin{aligned} & \left[ \psi_{\setminus \tau^{sum(\overline{\alpha}_m)}} if_{1,IF_{r,r,\tau(\cdot),\bar{i}_{r,s}}^{(s)}}^{(s)}(\cdot) (O_{i_{m+1}}; \theta) \right. \\ & \quad \times \prod_{(r_1,s_1) \neq (r,s)} IF_{r_1,r_1,\tau(\cdot),\bar{i}_{r_1,s_1}}^{(s)}(\theta) \left. \right] \\ & + \left[ \frac{\partial^{sum(\overline{\alpha}_m)} \left( if_{1,IF_{r,r,\psi(\tau,\cdot),\bar{i}_{r,s}}^{(s)}}^{(s)}(\cdot) (O_{i_{m+1}}; \theta) \right)}{\partial \tau^{[sum(\overline{\alpha}_m)-1]}} \right. \\ & \quad \times \prod_{(r_1,s_1) \neq (r,s)} IF_{r_1,r_1,\tau(\cdot),\bar{i}_{r_1,s_1}}^{(s)}(\theta) \left. \right] \\ & + \left( \sum_{(r_1,s_1) \neq (r,s)} \left[ \frac{\partial^{[sum(\overline{\alpha}_m)-1]} \left( IF_{r_1,r_1,\psi(\tau,\cdot),\bar{i}_{r_1,s_1}}^{(s)} \right)}{\partial \tau^{[sum(\overline{\alpha}_m)-1]}} \times \right. \right. \\ & \quad if_{1,IF_{r,r,\tau(\cdot),\bar{i}_{r,s}}^{(s)}}^{(s)}(\cdot) (O_{i_{m+1}}; \theta) \times \left. \prod_{(r_2,s_2) \neq (r_1,s_1) \neq (r,s)} IF_{r_2,r_2,\tau(\cdot),\bar{i}_{r_2,s_2}}^{(s)}(\theta) \right] \left. \right) \end{aligned} \right)$$

Now, we have shown that, in addition to  $if_{1,if_{m,m,\psi(\tau,\cdot)}(\mathbf{o}_{i_m};\cdot)}^{(s)}(O_{i_{m+1}};\theta)$ , the RHS of eq.(75) can be written as the sum of three pieces with the following leading terms

$$\begin{aligned}
& \psi_{\setminus \tau^2} IF_{1,\tau(\cdot),i_{m+1}}(\theta) if_{m,m,\tau(\cdot)}^{(s)}(\mathbf{o}_{i_m};\theta) \\
& + \sum_{(\varkappa_{m,1},\varkappa_{m,2},\dots,\varkappa_{m,m-1})} \left( \prod_{r=1}^{m-1} \varkappa_{m,r}! \right)^{-1} \times \\
& \left\{ \left[ \psi_{\setminus \tau^{sum(\bar{\varkappa}_m)+1}} IF_{1,\tau(\cdot),i_{m+1}}(\theta) \right] \prod_{r=1}^{m-1} \left( \prod_{s=1}^{\varkappa_{m,r}} IF_{r,r,\tau(\cdot),\bar{i}_{r,s}}^{(s)}(\theta) \right) \right. \\
& \left. + \sum_{(r,s)}^{sum(\bar{\varkappa}_m)} \left( \psi_{\setminus \tau^{sum(\bar{\varkappa}_m)}} if_{1,IF_{r,r,\tau(\cdot),\bar{i}_{r,s}}^{(s)}(\cdot)}^{(s)}(O_{i_{m+1}};\theta) \right) \right. \\
& \left. \times \prod_{(r_1,s_1) \neq (r,s)} IF_{r_1,r_1,\tau(\cdot),\bar{i}_{r_1,s_1}}^{(s)}(\theta) \right) \left. \right\}, \quad (76)
\end{aligned}$$

while the remaining terms can be constructed by applying the algorithm **TE** to the above leading terms.

By part 5c) of Theorem 3 and the induction assumption, next, we only need to prove that eq. (76) is actually a kernel of the following  $(m+1)$ th order U-statistic

$$\begin{aligned}
& (m+1) \sum_{(\varkappa_{m+1,1},\varkappa_{m+1,2},\dots,\varkappa_{m+1,m})} \left( \prod_{r=1}^m \varkappa_{m+1,r}! \right)^{-1} \times \\
& \mathbb{V} \left\{ \psi_{\setminus \tau^{sum(\bar{\varkappa}_{m+1})}} \prod_{r=1}^m \left( \prod_{s=1}^{\varkappa_{m+1,r}} IF_{r,r,\tau(\cdot),\bar{i}_{r,s}^*}^{(s)}(\theta) \right) \right\} \quad (77)
\end{aligned}$$



where

$$\begin{aligned}\bar{\mathcal{X}}_{m+1} &\equiv (\mathcal{X}_{m+1,1}, \mathcal{X}_{m+1,2}, \dots, \mathcal{X}_{m+1,m}), \\ \text{and } m+1 &= \sum_{r=1}^m \mathcal{X}_{m+1,r} \times r, \quad \mathcal{X}_{m+1,r} \geq 0 \quad \forall 1 \leq r \leq m \\ \text{sum}(\bar{\mathcal{X}}_{m+1}) &\equiv \sum_{r=1}^m \mathcal{X}_{m+1,r} \\ \bar{l}_{r,s}^* &\equiv (l_{l_{r,s}^*+1}^*, \dots, l_{l_{r,s}^*+r}^*) \quad \text{where } l_{r,s}^* \equiv \sum_{q=1}^{r-1} \mathcal{X}_{m+1,q} \times q + (s-1)r\end{aligned}$$

This can be proved following a simple but important fact that, for any sum representation of  $m+1 = \sum_{r=1}^m \mathcal{X}_{m+1,r} \times r$ , either i)  $\mathcal{X}_{m+1,1} = \mathcal{X}_{m+1,m} = 1$ , and  $\mathcal{X}_{m+1,r} = 0$  for  $\forall 1 < r < m$ ; or ii)  $\mathcal{X}_{m+1,m} = 0$  and there exists a sum representation of  $m = \sum_{r=1}^{m-1} \mathcal{X}_{m,r} \times r$  such that,

iia)  $\mathcal{X}_{m+1,1} = \mathcal{X}_{m,1} + 1$  and  $\mathcal{X}_{m+1,r} = \mathcal{X}_{m,r}$  for  $1 < r < m$ , or

iib)  $\exists 1 \leq r^* \leq m-2$ , such that  $\mathcal{X}_{m+1,r^*} = \mathcal{X}_{m,r^*} - 1$ ,  $\mathcal{X}_{m+1,r^*+1} = \mathcal{X}_{m,r^*+1} + 1$ , and  $\mathcal{X}_{m+1,r} = \mathcal{X}_{m,r}$  for  $\forall r \neq r^*, r^*+1$ .

Define

$$\begin{aligned}
& LT1(\varkappa_{m,1}, \dots, \varkappa_{m,m-1}) \\
&= \left( \prod_{r=1}^{m-1} \varkappa_{m,r}! \right)^{-1} \left[ \psi_{\setminus \tau^{sum}(\bar{\varkappa}_m)+1} IF_{1,\tau(\cdot),i_{m+1}}(\theta) \right] \\
&\times \prod_{r=1}^{m-1} \left( \prod_{s=1}^{\varkappa_{m,r}} IF_{r,r,\tau(\cdot),\bar{i}_{r,s}}^{(s)}(\theta) \right) \\
& LT2(\varkappa_{m,1}, \dots, \varkappa_{m,m-1}; r) \\
&= \left( \prod_{r=1}^{m-1} \varkappa_{m,r}! \right)^{-1} \sum_{s=1}^{\varkappa_{m,r}} \left( \psi_{\setminus \tau^{sum}(\bar{\varkappa}_m)} if_{1,IF_{r,r,\tau(\cdot),\bar{i}_{r,s}}^{(s)}(\cdot)}(O_{i_{m+1}}; \theta) \right. \\
&\quad \times \left. \prod_{(r_1,s_1) \neq (r,s)} IF_{r_1,r_1,\tau(\cdot),\bar{i}_{r_1,s_1}}^{(s)}(\theta) \right).
\end{aligned}$$

Note that for any  $(\varkappa_{m+1,1}, \varkappa_{m+1,2}, \dots, \varkappa_{m+1,m} = 0)$ , if  $\varkappa_{m+1,1} > 0$ , then

$$\begin{aligned}
& \mathbb{V} \{ LT1(\varkappa_{m+1,1} - 1, \varkappa_{m+1,2}, \dots, \varkappa_{m+1,m-1}) \} \\
&= \varkappa_{m+1,1} \left( \prod_{r=1}^m \varkappa_{m+1,r}! \right)^{-1} \mathbb{V} \left\{ \psi_{\setminus \tau^{sum}(\bar{\varkappa}_{m+1})} \prod_{r=1}^m \left( \prod_{s=1}^{\varkappa_{m+1,r}} IF_{r,r,\tau(\cdot),\bar{i}_{r,s}}^{(s)}(\theta) \right) \right\},
\end{aligned}$$

and for any  $2 \leq r \leq m-1$  such that  $\varkappa_{m+1,r} > 0$ ,

$$\begin{aligned}
& \mathbb{V} \{ d_{m+1}(LT2(\varkappa_{m+1,1}, \dots, \varkappa_{m+1,r-1} + 1, \varkappa_{m+1,r} - 1, \dots, \varkappa_{m+1,m-1}, r)) \} \\
&= \varkappa_{m+1,r} \left( \prod_{r=1}^m \varkappa_{m+1,r}! \right)^{-1} \mathbb{V} \left\{ \psi_{\setminus \tau^{sum}(\bar{\varkappa}_{m+1})} \prod_{r=1}^m \left( \prod_{s=1}^{\varkappa_{m+1,r}} IF_{r,r,\tau(\cdot),\bar{i}_{r,s}^*}^{(s)}(\theta) \right) \right\}.
\end{aligned}$$

As  $\sum_{r=1}^{m-1} r \varkappa_{m+1,r} = m+1$ . Now, it is obvious that the term with  $(\varkappa_{m+1,1}, \varkappa_{m+1,2}, \dots, \varkappa_{m+1,m} = 0)$  in eq. (77) comes from the following terms in eq. (76)

$$I \{ \varkappa_{m+1,1} > 0 \} LT1 (\varkappa_{m+1,1} - 1, \varkappa_{m+1,2}, \dots, \varkappa_{m+1,m-1}) \\ + \sum_{r=2}^{m-1} I \{ \varkappa_{m+1,r} > 0 \} LT2 (\varkappa_{m+1,1}, \dots, \varkappa_{m+1,r-1} + 1, \varkappa_{m+1,r} - 1, \dots, \varkappa_{m+1,m-1}, r),$$

while the term with  $(\varkappa_{m+1,1} = 1, 0, \dots, 0, \varkappa_{m+1,m} = 1)$  in eq. (77) comes from the following terms in eq. (76)

$$\psi_{\setminus \tau^2} IF_{1, \tau(\cdot), i_{m+1}}(\theta) if_{m, m, \tau(\cdot)}^{(s)}(\mathbf{o}_{i_m}; \theta) \\ + \psi_{\setminus \tau^2} if_{1, IF_{m-1, m-1, \tau(\cdot), \bar{i}_{m-1, 1}}^{(s)}}(\cdot) (O_{i_{m+1}}; \theta) IF_{1, \tau(\cdot), i_1}^{(s)}(\theta).$$

■

**Proof.** (Theorem 45) We proceed by induction. For  $j = 1$ ,

$$\psi_{\setminus l_1}(\theta) = \sum_{r=1}^{\zeta} \left\{ \psi_{r, \setminus l_1}(\theta) \times \left[ \prod_{s \leq \zeta, s \neq r} \psi_s(\theta) \right] \right\} \\ = \sum_{r=1}^{\zeta} \left\{ E_{\theta} [\mathbb{IF}_{\psi_r, 1}(\theta) \zeta_{1; \setminus l_1}(\theta)] \times \left[ \prod_{s \leq \zeta, s \neq r} \psi_s(\theta) \right] \right\}$$

therefore

$$\mathbb{IF}_{\psi(\theta; \zeta), 1, 1}(\theta) = \mathbb{V} \left\{ \sum_{r=1}^{\zeta} IF_{\psi_r, 1, 1; i_1}(\theta) \times \left[ \prod_{s \leq \zeta, s \neq r} \psi_s(\theta) \right] \right\} \\ = \mathbb{V} \left[ \sum_{\{t_1, \dots, t_{\zeta}\} \in \Upsilon_{\zeta, 1}} \prod_{s=1}^{\zeta} IF_{\psi_s(\theta); t_s, t_s; \bar{i}_s, t_s}(\theta) \right]$$

Assume that the lemma holds for  $j$ , i.e.:

$$\mathbb{IF}_{\psi(\theta;\zeta);j,j}(\theta) = \mathbb{V} \left[ \sum_{\{t_1, \dots, t_\zeta\} \in \Upsilon_{\zeta;j}} \prod_{s=1}^{\zeta} IF_{\psi_s(\theta);t_s,t_s;\bar{i}_s,t_s}(\theta) \right]$$

we now show that it holds for  $j+1$ . Now,

$$\begin{aligned} & (j+1) \mathbb{IF}_{\psi(\theta;\zeta);j+1,j+1}(\theta) \\ &= \mathbb{V} \left[ if_{if_{\psi(\theta;\zeta);j,j}(O_{i_1}, \dots, O_{i_j}, \cdot);1,1} (O_{i_{j+1}}; \theta) \right] \\ &= \Pi_{\theta,m} \left[ \mathbb{V} \left[ if_{1,if_{\psi(\theta;\zeta);j,j}(O_{i_1}, \dots, O_{i_j}, \cdot)} (O_{i_{j+1}}; \theta) \right] | \mathcal{U}_j(\theta) \right] \end{aligned}$$

so that  $(j+1) \mathbb{IF}_{\psi(\theta;\zeta);j+1,j+1}(\theta) =$

$$\begin{aligned} & \mathbb{V} \left[ \sum_{\{t_1, \dots, t_\zeta\} \in \Upsilon_{\zeta;j}} \sum_{r=1}^{\zeta} \left\{ \left\{ \begin{aligned} & if_{IF_{\psi_r(\theta);t_r,t_r;\bar{i}_r,t_r}(\theta);1,1} (O_{i_{r,t_r+1}}; \theta) \\ & - \Pi_{\theta,t_r} \left[ \mathbb{V} \left[ if_{IF_{\psi_r(\theta);t_r,t_r;\bar{i}_r,t_r}(\theta);1,1} (O_{i_{r,t_r+1}}; \theta) \right] | \mathcal{U}_{t_r}(\theta) \right] \\ & \times \left( \prod_{s \leq \zeta, s \neq r} IF_{\psi_s(\theta);t_s,t_s;\bar{i}_s,t_s}(\theta) \right) \end{aligned} \right\} \right\} \right] \\ &= \mathbb{V} \left[ \sum_{\{t_1, \dots, t_\zeta\} \in \Upsilon_{\zeta;j}} \sum_{r=1}^{\zeta} \left\{ \begin{aligned} & (t_r+1) IF_{\psi_r(\theta);t_r+1,t_r+1;\bar{i}_r,t_r+1}(\theta) \\ & \times \left( \prod_{s \leq \zeta, s \neq r} IF_{\psi_s(\theta);t_s,t_s;\bar{i}_s,t_s}(\theta) \right) \end{aligned} \right\} \right] \end{aligned}$$

Now, consider an arbitrary term in the double sum corresponding to the index vector

$$(t'_1, \dots, t'_{r^*}, \dots, t'_\zeta)$$

where the star indicates the index of the second summation . Then this term and terms corresponding to

$$\begin{aligned}
& (t'_{1*} - 1, t'_2, \dots, t'_r + 1, \dots, t'_\zeta), \\
& (t'_1, t'_{2*} - 1, \dots, t'_r + 1, \dots, t'_\zeta), \\
& \vdots \\
& (t'_1, \dots, t'_{r-1*} - 1, t'_r + 1, \dots, t'_\zeta), \\
& (t'_1, \dots, t'_r + 1, t'_{r+1*} - 1, \dots, t'_\zeta), \\
& \vdots \\
& (t'_1, \dots, t'_r + 1, \dots, t'_\zeta - 1)
\end{aligned}$$

all share the common factor

$$IF_{\psi_{r*}(\theta); t_{r*}+1, t_{r*}+1; \bar{i}_{r*}, t_{r*}+1}(\theta) \times \left( \prod_{s \leq \zeta, s \neq r*} IF_{\psi_s(\theta); t'_s, t'_s; \bar{i}_{s, t'_s}}(\theta) \right)$$

but with respective multiplicative constants  $(t'_{r*} + 1), t'_{1*}, t'_{2*}, \dots, t'_{r-1*}, t'_{r+1*}, \dots, t'_\zeta$ . So that the total contribution of terms in the double summation that share this common factor is given by:

$$\begin{aligned}
& \left( (t'_{r*} + 1) + \sum_{r \neq r*} t'_r \right) \left( IF_{\psi_{r*}(\theta); t'_{r*}+1, t'_{r*}+1; \bar{i}_{r*}, t'_{r*}+1}(\theta) \right. \\
& \quad \left. \times \left( \prod_{s \leq \zeta, s \neq r*} IF_{\psi_s(\theta); t'_s, t'_s; \bar{i}_{s, t'_s}}(\theta) \right) \right) \\
& = (j + 1) \left( IF_{\psi_{r*}(\theta); t'_{r*}+1, t'_{r*}+1; \bar{i}_{r*}, t'_{r*}+1}(\theta) \right. \\
& \quad \left. \times \left( \prod_{s \leq \zeta, s \neq r*} IF_{\psi_s(\theta); t'_s, t'_s; \bar{i}_{s, t'_s}}(\theta) \right) \right)
\end{aligned}$$

Repeating this argument over all common factors in the set

$$\mathcal{A} = \left\{ \left( \left( IF_{\psi_{r^*}(\theta); t_{r^*}+1, t_{r^*}+1; \bar{t}_{r^*}, t_{r^*}+1}(\theta) \right) \times \left( \prod_{s \leq \zeta, s \neq r^*} IF_{\psi_s(\theta); t'_s, t'_s; \bar{t}'_s, t'_s}(\theta) \right) \right) : \{t_1, \dots, t_\zeta\} \in \Upsilon_{\zeta; j} \right\}$$

that appear in the double sum, we recover the desired sum

$$(j+1) \mathbb{I}\mathbb{F}_{\psi(\theta; \zeta); j+1, j+1}(\theta) = (j+1) \mathbb{V} \left[ \sum_{\{t_1, \dots, t_\zeta\} \in \Upsilon_{\zeta; j+1}} \left( \prod_{s \leq \zeta} IF_{\psi_s(\theta); t_s, t_s; \bar{t}_s, t_s}(\theta) \right) \right]$$

This is because the set of all common factors in  $\mathcal{A}$  is precisely :

$$\left\{ \prod_{s \leq \zeta} IF_{\psi_s(\theta); t_s, t_s; \bar{t}_s, t_s}(\theta) : \{t_1, \dots, t_\zeta\} \in \Upsilon_{\zeta; j+1} \right\}$$

This concludes the proof. ■

**Proof.** (Theorem 49) for  $m = 1$ ,

$$\begin{aligned} & E_\theta \left( \widehat{\psi}_1 \right) - \psi(\theta) \\ &= E_\theta \left[ \frac{R_1}{\widehat{\pi}_1} \frac{R_0}{\widehat{\pi}_0} \left( Y - \widehat{B}_1 \right) \right] + E_\theta \left[ \frac{R_0}{\widehat{\pi}_0} \left( \widehat{B}_1 - \widehat{B}_0 \right) \right] + E_\theta \left( \widehat{B}_0 - B_0 \right) \\ &= E_\theta \left[ \frac{R_0}{\widehat{\pi}_0} \delta P_1 \delta B_1 + \delta P_0 \delta B_0 \right] \end{aligned}$$

Next we proceed to prove eq.(56) and eq.(59) by induction,

First of all,

$$\begin{aligned}
& E_\theta \left[ \frac{R_1}{\widehat{\pi}_1} \frac{R_0}{\widehat{\pi}_0} (Y - \widehat{B}_1) + \frac{R_0}{\widehat{\pi}_0} (\widehat{B}_1 - \widehat{B}_0) | L_0 \right] \\
&= E_\theta \left( \frac{R_0}{\widehat{\pi}_0} (B_1 - \widehat{B}_1) \left( \frac{R_1}{\widehat{\pi}_1} - 1 \right) + \frac{R_0}{\widehat{\pi}_0} (B_1 - \widehat{B}_1 + \widehat{B}_1 - \widehat{B}_0) | L_0 \right) \\
&= E_\theta \left( \frac{R_0}{\widehat{\pi}_0} \delta B_1 \delta P_1 + \frac{R_0}{\widehat{\pi}_0} \delta B_0 | L_0 \right)
\end{aligned}$$

and

$$\begin{aligned}
& E_\theta \left[ \frac{R_0}{\widehat{\pi}_0} \delta P_1 \delta B_1 \right] \\
&= E_\theta \left( \Pi_\theta \left( q_{01}^{1/2} \delta B_1 | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right) \Pi_\theta \left( q_0 q_{01}^{-1/2} \delta P_1 | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right) \right) + \\
& E_\theta \left( \Pi_\theta^\perp \left( q_{01}^{1/2} \delta B_1 | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right) \Pi_\theta^\perp \left( q_0 q_{01}^{-1/2} \delta P_1 | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right) \right)
\end{aligned}$$

with

$$\begin{aligned}
& E_\theta \left( \Pi_\theta \left( q_{01}^{1/2} \delta B_1 | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right) \Pi_\theta \left( q_0 q_{01}^{-1/2} \delta P_1 | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right) \right) \\
&= E_\theta \left( q_{01} \delta B_1 \overline{W}_{k_1}^T \right) E_\theta \left( q_{01} \overline{W}_{k_1} \overline{W}_{k_1}^T \right)^{-1} E_\theta \left( q_0 \delta P_1 \overline{W}_{k_1} \right)
\end{aligned}$$

For  $m = 2$ ,

$$\begin{aligned}
& E_\theta \left( \widehat{\psi}_2 \right) - \psi(\theta) \\
&= E_\theta \left[ \frac{R_0}{\widehat{\pi}_0} \delta P_1 \delta B_1 + \delta P_0 \delta B_0 \right] + E_\theta \left( \widehat{\psi}_{2,2} \right) \\
&= E_\theta \left[ \frac{R_0}{\widehat{\pi}_0} \delta P_1 \delta B_1 + \delta P_0 \delta B_0 \right] - E_\theta \left[ \frac{R_0 \pi_1}{\widehat{\pi}_0 \widehat{\pi}_1} \delta B_1 \overline{W}_{k_1}^T \right] E_\theta \left[ \overline{W}_{k_1} \frac{R_0}{\widehat{\pi}_0} \delta P_1 \right] \\
&- E_\theta \left\{ \left( \frac{R_0}{\widehat{\pi}_0} \delta P_1 \delta B_1 + \frac{R_0}{\widehat{\pi}_0} \delta B_0 \right) \overline{Z}_{k_0}^T \right\} E_\theta \left[ \overline{Z}_{k_0} \delta P_0 \right] \\
&= -E_\theta \left( q_0 \delta B_0 \overline{Z}_{k_0}^T \right) E_\theta \left( q_0 \overline{Z}_{k_0} \overline{Z}_{k_0}^T \right)^{-1} \left[ E_\theta \left( q_0 \overline{Z}_{k_0} \overline{Z}_{k_0}^T \right) - I \right] E_\theta \left( \overline{Z}_{k_0} \delta P_0 \right) \\
&- E_\theta \left( q_{01} \delta B_1 \overline{W}_{k_1}^T \right) E_\theta \left( q_{01} \overline{W}_{k_1} \overline{W}_{k_1}^T \right)^{-1} \left( E_\theta \left( q_{01} \overline{W}_{k_1} \overline{W}_{k_1}^T \right) - I \right) E_\theta \left( \overline{W}_{k_1} q_0 \delta P_1 \right) \\
&- E_\theta \left( q_0 \delta P_1 \delta B_1 \overline{Z}_{k_0}^T \right) E_\theta \left[ \overline{Z}_{k_0} \delta P_0 \right] \\
&+ E_\theta \left( \Pi_\theta^\perp \left( q_0^{1/2} \delta B_0 \mid \left( q_0^{1/2} \overline{Z}_{k_0} \right) \right) \Pi_\theta^\perp \left( q_0^{-1/2} \delta P_0 \mid \left( q_0^{1/2} \overline{Z}_{k_0} \right) \right) \right) \\
&+ E_\theta \left( \Pi_\theta^\perp \left( q_{01}^{1/2} \delta B_1 \mid \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right) \Pi_\theta^\perp \left( q_0 q_{01}^{-1/2} \delta P_1 \mid \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right) \right)
\end{aligned}$$



If e.q(56) holds for  $m - 1 \geq 2$ , we next show it also holds for  $m$ ,

$$\begin{aligned}
& E_\theta \left( \widehat{\psi}_m \right) - \psi(\theta) \\
&= E_\theta \left( \widehat{\psi}_{m,m} \right) + \left( E_\theta \left( \widehat{\psi}_{m-1} \right) - \widehat{\psi}_m \right) \\
&= BI_{m-1,2} + (-1)^{m-1} \times \\
& E_\theta \left\{ \begin{aligned} & E_\theta \left[ q_0 \delta P_1 \delta B_1 \bar{Z}_{k_0}^T \right] \left[ E_\theta \left( q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T - I \right) \right]^{m-2} E_\theta \left( \bar{Z}_{k_0} \delta P_0 \right) \\ & + E_\theta \left[ q_0 \delta B_0 \bar{Z}_{k_0}^T \right] \left[ E_\theta \left( q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T - I \right) \right]^{m-2} E_\theta \left( \bar{Z}_{k_0} \delta P_0 \right)_{-(L1.1)} \\ & + E_\theta \left( q_{01} \delta B_1 \bar{W}_{k_1}^T \right) \left[ E_\theta \left( q_{01} \bar{W}_{k_1} \bar{W}_{k_1}^T - I \right) \right]^{m-2} E_\theta \left( \bar{W}_{k_1} q_0 \delta P_1 \right)_{-(L1.2)} \\ & + \sum_{j=2}^{m-1} E_\theta \left[ \bar{Z}_{k_0} \delta P_0^T \right] \left[ E_\theta \left( q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T - I \right) \right]^{j-2} E_\theta \left[ q_{01} \delta B_1 \bar{Z}_{k_0} \bar{W}_{k_1}^T \right] \\ & \quad \times E_\theta \left( q_{01} \bar{W}_{k_1} \bar{W}_{k_1}^T - I \right)^{m-1-j} E_\theta \left[ \bar{W}_{k_1} q_0 \delta P_1 \right]_{-(L1.3.j)} \end{aligned} \right\} \\
& \quad - (-1)^{m-1} \times \\
& \left\{ \begin{aligned} & \left\{ E_\theta \left\{ [q_0 \delta B_0] \bar{Z}_{k_0}^T \right\} E_\theta \left[ q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T \right]^{-1} \times \right. \\ & \quad \left. \left[ E_\theta \left( q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T \right) - I \right]^{m-2} E_\theta \left[ \bar{Z}_{k_0} \delta P_0 \right] \right\}_{-(L2.1)} \\ & + \left\{ E_\theta \left\{ [q_{01} \delta B_1] \bar{W}_{k_1}^T \right\} E_\theta \left[ q_{01} \bar{W}_{k_1} \bar{W}_{k_1}^T \right]^{-1} \times \right. \\ & \quad \left. \left[ E_\theta \left( q_{01} \bar{W}_{k_1} \bar{W}_{k_1}^T \right) - I \right]^{m-2} E_\theta \left[ \bar{W}_{k_1} q_0 \delta P_1 \right] \right\}_{-(L2.2)} \\ & + \sum_{j=2}^{m-2} \left\{ E_\theta \left[ \bar{Z}_{k_0} \delta P_0^T \right] \left[ E_\theta \left( q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T - I \right) \right]^{j-2} \times \right. \\ & \quad E_\theta \left[ q_{01} \delta B_1 \bar{Z}_{k_0} \bar{W}_{k_1}^T \right] E_\theta \left[ q_{01} \bar{W}_{k_1} \bar{W}_{k_1}^T \right]^{-1} \times \\ & \quad \left. E_\theta \left( q_{01} \bar{W}_{k_1} \bar{W}_{k_1}^T - I \right)^{m-1-j} E_\theta \left[ \bar{W}_{k_1} q_0 \delta P_1 \right] \right\}_{-(L2.3.j)} \\ & + E_\theta \left\{ [q_0 \delta P_1 \delta B_1] \bar{Z}_{k_0}^T \right\} \left[ E_\theta \left( q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T \right) - I \right]^{m-3} E_\theta \left[ \bar{Z}_{k_0} \delta P_0 \right]_{-(L2.4)} \end{aligned} \right\}
\end{aligned}$$

It can be shown

$$\begin{aligned}
& (L1.1) - (L2.1) \\
&= (-1)^{m-1} \left\{ \begin{array}{l} E_{\theta} \left\{ [q_0 \delta B_0] \bar{Z}_{k_0}^T \right\} E_{\theta} \left[ q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T \right]^{-1} \times \\ \left[ E_{\theta} \left( q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T \right) - I \right]^{m-1} E_{\theta} \left[ \bar{Z}_{k_0} \delta P_0 \right] \end{array} \right\}
\end{aligned}$$

$$\begin{aligned}
& (L1.2) - (L2.2) \\
&= (-1)^{m-1} \left\{ \begin{array}{l} E_{\theta} \left\{ [q_{01} \delta B_1] \bar{W}_{k_1}^T \right\} E_{\theta} \left[ q_{01} \bar{W}_{k_1} \bar{W}_{k_1}^T \right]^{-1} \times \\ \left[ E_{\theta} \left( q_{01} \bar{W}_{k_1} \bar{W}_{k_1}^T \right) - I \right]^{m-1} E_{\theta} \left[ \bar{W}_{k_1} q_0 \delta P_1 \right] \end{array} \right\}
\end{aligned}$$

$$\forall 2 \leq j < m-1,$$

$$\begin{aligned}
& (L1.3.j) - (L2.3.j) \\
&= (-1)^{m-1} \left\{ \begin{array}{l} E_{\theta} \left[ \bar{Z}_{k_0} \delta P_0^T \right] \left[ E_{\theta} \left( q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T - I \right) \right]^{j-2} \times \\ E_{\theta} \left[ q_{01} \delta B_1 \bar{Z}_{k_0} \bar{W}_{k_1}^T \right] E_{\theta} \left[ q_{01} \bar{W}_{k_1} \bar{W}_{k_1}^T \right]^{-1} \times \\ E_{\theta} \left( q_{01} \bar{W}_{k_1} \bar{W}_{k_1}^T - I \right)^{m-j} E_{\theta} \left[ \bar{W}_{k_1} q_0 \delta P_1 \right] \end{array} \right\}
\end{aligned}$$

If  $\zeta_m(L_0, \theta) \equiv E_\theta [\delta P_0 \bar{Z}_{k_0}^T] [E_\theta (q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T - I)]^{m-3} \bar{Z}_{k_0}$ , then

$$\begin{aligned}
& (L1.3.m-1) - (L2.4) \\
&= E_\theta [q_{01} \delta B_1 \zeta_m(L_0, \theta) \bar{W}_{k_1}^T] E_\theta [\bar{W}_{k_1} q_0 \delta P_1] - E_\theta \{\zeta_m(L_0, \theta) q_0 \delta P_1 \delta B_1\} \\
&= (-1)^{m-1} \left\{ \begin{aligned} & E_\theta (q_{01} \delta B_1 \zeta_m(L_0, \theta) \bar{W}_{k_1}^T) E_\theta (q_{01} \bar{W}_{k_1} \bar{W}_{k_1}^T)^{-1} \times \\ & (E_\theta (q_{01} \bar{W}_{k_1} \bar{W}_{k_1}^T - I)) E_\theta (\bar{W}_{k_1} q_0 \delta P_1) \end{aligned} \right\} \\
&+ (-1)^m E_\theta \left( \begin{aligned} & \Pi_\theta^\perp \left[ \left( q_{01}^{1/2} \delta B_1 \zeta_m(L_0, \theta) \right) | \left( q_{01}^{1/2} \bar{W}_{k_1} \right) \right] \\ & \times \Pi_\theta^\perp \left[ q_0 q_{01}^{-1/2} \delta P_1 | \left( q_{01}^{1/2} \bar{W}_{k_1} \right) \right] \end{aligned} \right)_{-(TB_{m-1, m-1}^{(2)})}
\end{aligned}$$

In summary

$$\begin{aligned}
& E_\theta (\hat{\psi}_m) - \psi(\theta) \\
&= (-1)^{m-1} BI_{m,1} + BI_{m-1,2} + (TB_{m-1, m-1}^{(2)})
\end{aligned}$$

Next we show  $\left(TB_{m-1,m-1}^{(2)}\right) = BI_{m,2} - BI_{m-1,2}$ ,

$$\begin{aligned}
& \zeta_m(L_0, \theta) \\
&= E_\theta \left[ \delta P_0 \bar{Z}_{k_0}^T \right] \left[ E_\theta \left( q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T - I \right) \right]^{m-3} \bar{Z}_{k_0} \\
&= \left\{ E_\theta \left[ \delta P_0 \bar{Z}_{k_0}^T \right] \left( \begin{pmatrix} E_\theta \left( q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T \right)^{-1} + \\ I - E_\theta \left( q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T \right)^{-1} \end{pmatrix} \right) \right\} \\
&\quad \times \left[ E_\theta \left( q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T - I \right) \right]^{m-3} \bar{Z}_{k_0} \Bigg\} \\
&= \left\{ E_\theta \left[ \delta P_0 \bar{Z}_{k_0}^T \right] E_\theta \left( q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T \right)^{-1} \left[ E_\theta \left( q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T - I \right) \right]^{m-3} \bar{Z}_{k_0} \right\} \\
&+ \left\{ E_\theta \left[ \delta P_0 \bar{Z}_{k_0}^T \right] E_\theta \left( q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T \right)^{-1} E_\theta \left( q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T - I \right) \right\} \\
&\quad \times \left[ E_\theta \left( q_0 \bar{Z}_{k_0} \bar{Z}_{k_0}^T - I \right) \right]^{m-3} \bar{Z}_{k_0} \Bigg\}
\end{aligned}$$

Applying the above expression of  $\zeta_m(L_0, \theta)$  to e.q(L3.1), we find that

$$\left(TB_{m-1,m-1}^{(2)}\right) + BI_{m-1,2} = BI_{m,2}$$

which completes the induction.

We want to mention that  $TB_{j,j}^{(2)}$  ( $2 \leq j < m$ ) equals  $\sum_{l=1}^{k_0^{j-1}} \left( \prod_{t=1}^j \tilde{\tau}_{l,t}^{(j)}(\theta) - \prod_{t=1}^j \tau_{l,t}^{(j)}(\theta) \right)$ ,  $BI_{2,2} = \left( \tilde{\psi}^\dagger(\theta) - \psi(\theta) \right) + \tilde{\tau}_{1,1}^{(1)}(\theta) - \tau_{1,1}^{(1)}(\theta)$ ,  $EB_1^{(1)} = E_\theta \left( \mathbb{IF}_{m, \tilde{\psi}^\dagger}(\hat{\theta}) + \psi(\hat{\theta}) \right) - \tilde{\psi}^\dagger(\theta)$ ,  $EB_1^{(2)} = E \left( \mathbb{IF}_{m, \tilde{\tau}_{1,1}^{(1)}(\theta)}(\hat{\theta}) \right) - \tilde{\tau}_{1,1}^{(1)}(\theta)$ , and  $EB_{jj}^{(2)} = E \left( \mathbb{IF}_{m, \tilde{\psi}_{jj}(\theta)}(\hat{\theta}) \right) - \tilde{\psi}_{jj}(\theta)$ . Therefore the results from this theorem is consistent with eq.(55). Technical details are not presented here.

Eq.(57) follows from eq.(56) similarly as in the proof of theorem 49 and eq.(58) can be derived from eq.(57) by our assumption of rate optimality of the initial estimator.

Next we prove eq.(60),

$$\begin{aligned}
& |BI_{2,2}| \\
& \leq \left| E \left( \Pi_{\theta}^{\perp} \left( q_0^{1/2} \delta B_0 | \left( q_0^{1/2} \overline{Z}_{k_0} \right) \right) \Pi_{\theta}^{\perp} \left( q_0^{-1/2} \delta P_0 | \left( q_0^{1/2} \overline{Z}_{k_0} \right) \right) \right) \right| \\
& + \left| E \left( \Pi_{\theta}^{\perp} \left( q_{01}^{1/2} \delta B_1 | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right) \Pi_{\theta}^{\perp} \left( q_0 q_{01}^{-1/2} \delta P_1 | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right) \right) \right| \\
& \leq \left\{ E \left( \Pi_{\theta}^{\perp} \left( q_0^{1/2} \delta B_0 | \left( q_0^{1/2} \overline{Z}_{k_0} \right) \right) \right)^2 \right\}^{1/2} \left\{ E \left( \Pi_{\theta}^{\perp} \left( q_0^{-1/2} \delta P_0 | \left( q_0^{1/2} \overline{Z}_{k_0} \right) \right) \right)^2 \right\}^{1/2} \\
& + \left\{ E \left( \Pi_{\theta}^{\perp} \left( q_{01}^{1/2} \delta B_1 | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right) \right)^2 \right\}^{1/2} \left\{ E \left( \Pi_{\theta}^{\perp} \left( q_0 q_{01}^{-1/2} \delta P_1 | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right) \right)^2 \right\}^{1/2} \\
& = O_p \left( \max \left[ k_0^{-(\beta_{b_0} + \beta_{\pi_0})/d_0}, k_0^{-(\beta_{b_1} + \beta_{\pi_1})/d_1} \right] \right)
\end{aligned}$$

(Cauchy-Shwartz inequality. The last equality can easily be shown as in the proof of theorem 31)

$$\begin{aligned}
& |(TB(m, 2))| \\
& = \left| E \left[ \begin{aligned} & \Pi_{\theta}^{\perp} \left[ \left( \frac{\pi_1^{1/2}}{\hat{\pi}_1^{1/2}} \delta B_1 \Pi_{\theta} \left( q_0^{1/2} \delta P_0 | \left( q_0^{1/2} \overline{Z}_{k_0} \right) \right) \right) | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right] \\ & \times \Pi_{\theta}^{\perp} \left[ q_0 q_{01}^{-1/2} \delta P_1 | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right] \end{aligned} \right] \right| \\
& \leq \left| E \left[ \begin{aligned} & \Pi_{\theta}^{\perp} \left[ \left( \delta B_1 q_{01}^{1/2} \delta P_0 \right) | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right] \\ & \times \Pi_{\theta}^{\perp} \left[ q_0 q_{01}^{-1/2} \delta P_1 | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right] \end{aligned} \right] \right| \\
& + \left| E \left[ \begin{aligned} & \Pi_{\theta}^{\perp} \left[ \left( \frac{\pi_1^{1/2}}{\hat{\pi}_1^{1/2}} \delta B_1 \Pi_{\theta}^{\perp} \left( q_0^{1/2} \delta P_0 | \left( q_0^{1/2} \overline{Z}_{k_0} \right) \right) \right) | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right] \\ & \times \Pi_{\theta}^{\perp} \left[ q_0 q_{01}^{-1/2} \delta P_1 | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right] \end{aligned} \right] \right| \\
& \leq O_p \left( \max \left( k_1^{-(\min(\beta_{\pi_0}, \beta_{b_1}) + \beta_{\pi_1})/d_1}, k_1^{-\beta_{\pi_1}/d_1} k_0^{-\beta_{\pi_0}/d_0} \left( \frac{\log n}{n} \right)^{\frac{\beta_{b_1}}{d_1 + \beta_{b_1}}} \right) \right)
\end{aligned}$$

The last inequality holds because

$$\begin{aligned}
& E \left( \Pi_{\theta}^{\perp} \left[ \left( \frac{\pi_1^{1/2}}{\widehat{\pi}_1^{1/2}} \delta B_1 \Pi_{\theta}^{\perp} \left( q_0^{1/2} \delta P_0 | \left( q_0^{1/2} \overline{Z}_{k_0} \right) \right) \right) | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right] \right)^2 \\
& \leq E \left( \frac{\pi_1^{1/2}}{\widehat{\pi}_1^{1/2}} \delta B_1 \Pi_{\theta}^{\perp} \left( q_0^{1/2} \delta P_0 | \left( q_0^{1/2} \overline{Z}_{k_0} \right) \right) \right)^2 \\
& \leq \left\| \frac{\pi_1^{1/2}}{\widehat{\pi}_1^{1/2}} \delta B_1 \right\|_{\infty}^2 E \left( \Pi_{\theta}^{\perp} \left( q_0^{1/2} \delta P_0 | \left( q_0^{1/2} \overline{Z}_{k_0} \right) \right) \right)^2
\end{aligned}$$

(The first inequality holds because the orthogonal projection operator has a norm no greater than 1)

$$\begin{aligned}
& |TB(m, 3)| \\
& = \left| E \left[ \Pi_{\theta}^{\perp} \left[ \left( \begin{aligned} & q_{01}^{1/2} \delta B_1 E \left[ \delta P_0 \overline{Z}_{k_0}^T \right] E \left[ q_0 \overline{Z}_{k_0} \overline{Z}_{k_0}^T \right]^{-1} \\ & \times \left( E \left[ q_0 \overline{Z}_{k_0} \overline{Z}_{k_0}^T - I \right] \right)^{m-2} \overline{Z}_{k_0} \end{aligned} \right) | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right] \right] \right| \\
& \quad \times \Pi_{\theta}^{\perp} \left[ q_0 q_{01}^{-1/2} \delta P_1 | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right] \\
& \leq \left\{ \left\{ E \left( \begin{aligned} & q_{01}^{1/2} \delta B_1 E \left[ \delta P_0 \overline{Z}_{k_0}^T \right] E \left[ q_0 \overline{Z}_{k_0} \overline{Z}_{k_0}^T \right]^{-1} \\ & \times \left( E \left[ q_0 \overline{Z}_{k_0} \overline{Z}_{k_0}^T - I \right] \right)^{m-2} \overline{Z}_{k_0} \end{aligned} \right)^2 \right\}^{1/2} \right\} \\
& \quad \times \left\{ E \left( \Pi_{\theta}^{\perp} \left[ q_0 q_{01}^{-1/2} \delta P_1 | \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right] \right)^2 \right\}^{1/2}
\end{aligned}$$

(Cauchy-Schwartz inequality and projection operator has operator norm of 1)

$$\begin{aligned}
& \leq \left\{ \left\| q_{01}^{1/2} \frac{f_1}{\hat{f}_1} \right\|_{\infty} \|\delta B_1\|_{\infty} \left\| q_0^{-1/2} \frac{\hat{f}_0}{f_0} \right\|_{\infty} \left\| \frac{f_0}{\pi_0 \hat{\pi}_0} \right\|_{\infty} \right. \\
& \quad \times \|\delta g_0\|_{\infty}^{m-2} \left\{ \int (\pi_0 - \hat{\pi}_0)^2 dL_0 \right\}^{1/2} \times \\
& \quad \left. \left\{ E \left( \Pi_{\theta}^{\perp} \left[ q_0 q_{01}^{-1/2} \delta P_1 \left( q_{01}^{1/2} \overline{W}_{k_1} \right) \right] \right)^2 \right\}^{1/2} \right\} \\
& = O_p \left( \left( \frac{\log n}{n} \right)^{-\frac{\beta_{b_1}}{d_1 + \beta_{b_1}} - \frac{(m-2)\beta_{g_0}}{d_0 + \beta_{g_0}}} n^{-\frac{\beta_{\pi_0}}{d_0 + \beta_{\pi_0}}} k_1^{-\beta_{\pi_1}/d_1} \right)
\end{aligned}$$

■